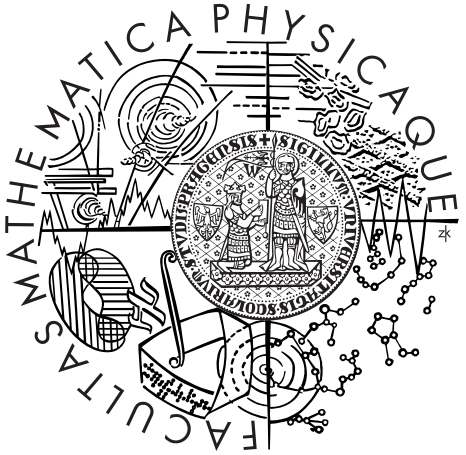
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Adrián Lachata

Automatické přiřazení diagnoz lékařským zprávám

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Jiří Hana, Ph.D.
Studijní program: Informatika
Studijní obor: Programování

Praha 2014

Ďakujem vedúcemu práce Jiřimu Hanovi za jeho záujem, profesionálne vedenie a ľudský prístup.

Ďakujem svojej rodine za jej neúnavnú celoživotnú podporu nie len pri písaní tejto práce a svojim priateľom za to, že mi robia život krajším.

V neposlednom rade ďakujem firme TERSINIDA a.s. za poskytnutie lekárskych správ spolu priradenými diagnózami pre účely tejto práce.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Automatické přiřazení diagnoz lékařským zprávám

Autor: Adrián Lachata

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: RNDr. Jiří Hana, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Cieľom práce bolo vyvinúť systém pre automatické priraďovanie kódov diagnóz českým lekárskym správam podľa medzinárodného číselníku diagnóz (ICD-10). Náš systém správy predspracováva a generuje atribúty podľa frekvencie, informačného zisku (IG) alebo vzájomnej informácie (PMI), na ktoré potom aplikuje klasifikačné algoritmy Naive Bayes a Rozhodovacie stromy. Skúčali sme viacero kombinácií. Rôžne počty atribútov, unigramy, bigramy a ich kombinácie ako aj rôzne formy predspracovania. Pri predspracovávaní sme využili morfólogický slovník pre normalizáciu textu, ignorovanie stopwords zo všeobecného zoznamu a vygenerovaných pomocou IDF. Atribúty sme filtrovali frekvencou, informačným ziskom a vzájomnou informáciou. Vybrali sme si pať diagnóz so zameraním na I10 (primárna hypertenze). Tie sme potom trénovali a testovali naším systémom na korpuse o jednom milióne lekárskych správ. Najsľubnejšie výsledky dosiahla diagnóza I10, na ktorej sme aj porovnali systém s človekom.

Klíčová slova: strojové učenie, kategorizácia textu, ICD–10

Title: Automatic assignment of diagnosis to medical reports

Author: Adrián Lachata

Department: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Jiří Hana, Ph.D., Institute of Formal and Applied Linguistics

Abstract: The goal of this thesis is to develop a system for automatic assignment of diagnosis codes (using the ICD-10 classification) to Czech textual medical reports. Our system preprocessed the text, generates key features, filters them to make their number manageable, and finally applies a supervised classification algorithm. At each step we have used and compared several alternatives: various counts of key features, unigrams and bigrams as feature types, stemming and morphology to normalize text words, ignoring general stopwords and stopwords generated by IDF, Information Gain and PMI metrics for feature space reduction and as classification algorithms Naive Bayes and Depedency Tries. As a pilot, we have selected 5 diagnoses, focusing mostly on the I10 diagnosis (Essential hypertension). For these 5 diagnoses, the system has been trained and tested on a corpus of one million medical reports. The most promising results are for I10. For I10, we have also compared our automatic results with classification performed manually by doctors of medicine.

Keywords: machine learning, text classification, ICD–10

Názov práce: Automatické přiřazení diagnoz lékařským zprávám

Autor: Adrián Lachata

Katedra: Ústav formální a aplikované lingvistiky

Vedúcí bakalárskej práce: RNDr. Jiří Hana, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Cieľom práce bolo vyvinúť systém pre automatické priraďovanie kódov diagnóz českým lekárskym správam podľa medzinárodného číselníku diagnóz (ICD-10). Náš systém správy predspracováva a generuje atribúty podľa frekvencie, informačného zisku (IG) alebo vzájomnej informácie (PMI), na ktoré potom aplikuje e klasifikačné algoritmy Naive Bayes a Rozhodovacie stromy. Skúčali sme viacero možností ako rôzne počty atribútov, unigramy a bigramy ako typy atribútov, stemming a morfólogický slovník pre normalizáciu textu, ignorovanie stopwords z všeobecného zoznamu a vygenerovaných pomocou IDF, informačný zisk a vzájomnú informáciu. Vybrali sme si pať diagnóz so zameraním na I10 (primárna hypertenze). Tie sme potom trénovali a testovali naším systémom na korpuse o jednom milióne lekárskych správ. Najsľubnejšie výsledky dosiahla diagnóza I10, na ktorej sme aj porovnali systém s človekom.

Kľúčové slová: strojové učenie, kategorizácia textu, ICD-10

Obsah

Úvod

Elektronické spracovanie dát je v dnešnom svete viac ako preferované. Preto niet divu, že lekárske ambulancie, polikliniky, nemocnice a iné zdravotnícke zariadenia elektronicky produkujú hromady dát. Medzi nimi aj lekárske správy spolu s priradenými diagnózami. My sme sa rozhodli preskúmať, či je možné na základe minulých správ priraďovať diagnózy automaticky.

Ak by to fungovalo, vidíme dve hlavné využitia. Predstavme si, že lekár píše správu o pacientovi, ktorého má pred sebou. Samozrejme, bežné diagnózy lekár nemá problém určiť sám. Ale nejaká vzácna diagnóza, ktorej správy majú podobný obsah ako správa, ktorú lekár píše, dotyčného lekára nemusí ani napadnúť. Nie preto, žeby lekár nebol profesionál, ale poznať všetky diagnózy je jednoducho nad ľudské sily. Program mu potom našepká zopár zriedkavých diagnóz, aby ich lekár vzal do úvahy a prípadne sa spýtal pacienta na doplňajúce informácie pre jednoznačné vylúčenie našepkaných diagnóz.

Druhé využitie je spätná kontrola. Máme lekárske správy s diagnózami, o ktorých vieme, že z nejakého dôvodu nie sú správne. Potom lekár môže pustiť počítačový program a zvážiť diagnózy, ktoré by program automaticky priradil.

Lekárske správy, ktoré máme k dispozícií, sú písané v českom jazyku, a pokiaľ nám je známe, nikto sa o podobnú vec pre český jazyk nepokúsil. Preto, so základnou znalosťou strojového učenia a spracovávania prirodzeného jazyka, sa pokúsime priradiť aspoň pár vybraných diagnóz.

1. Popis domény

Doménu dát tvorí dvojica lekárska správa – diagnóza. Dáta pochádzajú z rôznych ambulancií nachádzajúcich sa na území Českej. Správy sú písané v českom jazyku.

V tejto kapitole začneme príčinou vzniku lekárskych správ a uvedieme ich neexatnú definíciu. Následne sa zameriame na lekárske nálezy, ich štruktúru a vlastnosti. Potom predstavíme klasifikáciu chorôb spolu a v krátkosti spomenieme jej históriu. Ďalej popíšeme kapitoly klasifikácie chorôb. Nakoniec stručne kategorizujeme problém z informatického hľadiska.

1.1 Lekárske správy

Lekári môžu zasahovať do organizmu pacienta, a tým ho nejakým spôsobom ovplyvniť. Neskôr, pri návšteve iného, alebo aj toho istého lekára, môže lákar na základe znalostí z predchádzajúcich vyšetrení zvoliť iný postup liečby, ako keby nepoznal presnú pacientovú históriu. Poznáme viacero spôsobov zaznamenávania anamnézy a jedným z nich je práve lekárska správa. Ďalšie spôsoby sú napríklad laboratórne výsledky, rôzne merania, záznamy z operačného sálu a pod.

Bohužiaľ, presná definícia pojmu *lekárska správa* nám nie je známa a exaktný popis všetkých typov lekárskych správ nemá na výsledky tejto práce žiaden dopad, preto si dovolíme vlastné inštinktívne definície nasledujúcich pojmov. *Lekárska správa* je akýkoľvek dokument vyrobený lekárom, ktorý sa nejakým spôsobom vzťahuje k pacientovi. *Lekárský nález* je dokument vyrobený lekárom počas návštevy pacienta alebo bezprostredne po jeho odchode. Typicky sa jedná o súhrn objektívnych a subjektívnych príznakov pomáhajúcich pri určovaní diagnózy. Pre bližšie informácie viď *1.1.1* a *1.1.2*. Práve lekárske nálezy sú správy, ku ktorým budeme automaticky priraďovať diagnózy.

Ďalej v texte budeme používať pojem lekárska správa vo význame nami definovaného lekárskeho nálezu.

1.1.1 Štruktúra lekárskych správ

Lekárske správy nemajú svoju predpísanú štruktúru a sú písané formou voľného heslovitého textu. Avšak pri ich bližšom skúmaní je možné vypozorovať dve až tri zaužívané popisné časti nálezu, a to *subjektívna*, *objektívna* a *záver*. Subjektívna časť tradične obsahuje pacientove pocity a jeho laický pohľad na jeho zdravotný stav. V objektívnej časti lekár poskytne svoj odborný pohľad a záver zvykne obsahovať celkové zhrnutie a odporúčaný postup liečby.

1.1.2 Vlastnosti lekárskych správ

Text lekárskych správ má nasledujúce charakteristiky:

- Je písaný heslovite.
- Je krátky. Väčšina sa dĺžkou textu zmestí do piatich riadkov.
- Obsahuje veľa preklepov a gramatických chýb.

- Diakritika nie je pravidlom.
- Je nedôsledne formátovaný. Hlavne oddeľovanie slov.
- Obsahuje mnoho skrátených slov.
- Často mu chýba kontext.

Nedôsledné formátovanie, diakritika, preklepy a rôzne skrátené slová pre rovnaký termín sťažujú strojové spracovávanie, pretože slová s rovnakým významom alebo základným tvarom sú brané ako rozdielne. Posledná vlastnosť poukazuje na správy, ku ktorým bez predchádzajúceho kontextu nie je možné priradiť žiadnu diagnózu. Týka sa predovšetkým kontrol a nezmenenia zdravotného stavu od poslednej návštevy. Viď príklady 1.5 a 1.6.

K veľmi podobným vlastnostiam českých lekárskych správ došli aj [17] a [18].

Príklad 1.1. Subjektivní: Od pátku kašel, bez teplot, hleny nejdou. Kašel suchý. Brala Aerius - bez efektu. Trochu rýma. Bolesti v krku. Únava je. Dnes i torchu bolesti zad. V noci spí. Objektivní: Dýchání čisté sklípkové, bez vedl fenoménů. Závěr: Ad labor, KO + diff. Kontrola zítra zavolá.

Príklad 1.2. Kašel dlouzho - 2 meěsíce. Teploty 0- Rýma 0. Občas záděra na prstu. Objektivní: KP komp, dýchíní s expir pískoty. Hrdlo klidné. CRP méně než 8mg/l Závěr: Atrovent p.p. Erdomed.

Príklad 1.3. Subjektivní: Tel konsult: interpretacer výsledků - svědčí pro virosu. Objektivní: Již bez febrilií.

Príklad 1.4. Subjektivní: Bez akutních obtíží. Žádá očkování proti klíšť.encefalitidě. Aplikováno FSME i.m. č.š.VNR1H23D i.m., výkon bez komplikací. Samoplátce. Objektivní: Orientován, anemie 0, ikterus 0, bez cyanosy, AS pravidelná, bez šelestu. Břicho klidné, nebol. Hepart0 ,lien 0. Uzliny 0. Neurologicky bez lateralizace. EKG - nejasné ischem změny lat. Haemocult negat. Labor - hyperlipidemie, porucha glu toleravce. Závěr: Další prevence za 2 roky. Ad kardiologie. Za půl roku kontrola lipidogramu. Poučen o dietě.

Príklad 1.5. Subjektivní: Interpretace výlsdků, vše v pořádku.

Príklad 1.6. Subjektivní: Tel interpretace výsledků.

1.2 Klasifikácia chorôb

Klasifikácia chorôb sa dá definovať ako sústava kategórií, do ktorých sú zaraďované chorobné javy podľa zavedených kritérií. Urýchľuje domácu a medzinárodnú komunikáciu medzi lekármi, spätnú kontrolu, štatistické spracovanie, ako aj vedecké bádanie. Využívaná je aj zdravotnými poisťovňami pre bodové hodnotenie lekárov. [22]

1.2.1 Krátka história

Ľudstvo sa dlho klasifikáciou chorôb príliš nezaoberalo. Prvý pokus o systematickú klasifikáciu chôrob je známy až z 18. storočia a stojí za ním významný austrálsky štatistik známy ako *Sauvages*. Avšak v praxi sa začala používať až štatistická štúdia *Johna Graunta*, ktorá vznikla o storočie neskôr a východzala z údajov *London Bills of Mortality*. Graunt sa v nej pokúšal odhadnúť pomer úmrtia detí pred dosiahnutím šiesteho roku života, lebo údaje o veku neboli k dispozícií. Klasifikoval niekoľko pričín umŕtia, napríklad monoliázu, kŕče, nedostatok vitamínu D, kiahne alebo osípky. Napriek vysokej nepresnosti *Grautovej* štúdie, z neskorších dôkazov vyplýva, že v 36% prípadov boli jeho odhady správne. [24]

Užitočnosť klasifikácie chôrob vedúcich k úmrtiu rozpoznal *Prvý Medzinárodný Štatistký kongres (Brusel, 1853)* a požiadal *Williama Farra* a *Marca d'Espine* o vypracovanie medzinárodne akceptovateľnej a jednotnej klasifikácie pričín úmrtia. Obaja na ďalšiom kongrese predstavili dva nezávislé zoznamy založených na veľmi rozdielných princípoch. Farrov zoznam rozdeloval choroby do piatich kategórií. Epidémie, lokálne choroby podľa anatomického umiestenia, vývinové poruchy a choroby vzniknutú následkom násilia. D'Espine klasifikoval choroby podľa ich povahy. Kongres prijal kompromis, a hoci sa táto 1. revízia (revidovaná 1874, 1880 a 1886 Paríž) nikdy nestala medzinárodne akceptovateľnou, stala sa základom pre *Medzinárodný zoznam chorôb vedúcich k úmrtiu*. Zoznam sa postupne revidoval a až jeho šiesta revízia (New York, 1946) obsahovala aj iné ako choroby vedúce k úmrtiu. [24]

1.2.2 Súčasnosť

Aktuálna desiata revízia s českým názvom *Mezinárodní statistická klasifikace nemocí a přidružených zdravotních problémů (MKN–10)*, vo svete známeho ako *International Statistical Classification of Diseases and Related Health Problems (ICD–10)* v Českej Republike platí od roku 2004. Hlavným cieľom bolo poskytnúť pre každú chorobu jediný jednoznačný doporučený názov, ktorý by mal byť krátky, výstižný a založený na príčine. Celý proces koordinovala *Svetová Zdravotnícka Organizácia (WHO)*. [22]

1.3 Kapitoly klasifikácie chorôb

Klasifikačné choroby sa delia na kapitoly líšiace sa začiatočným písmenom. Konkrétne na:

- A,B – infekčné a parazitové choroby (A38 – šarlach)
- C – nádory (C089 – veľká slinná žľaza)
- D – choroby krvi a poruchy imunity
- E – choroby žliaz s vnútorným vylučovaním (E67.8 – nadmerná výživa)
- F – duševné poruchy a poruchy správania (F41.9 – úzkostná porucha bližšie neurčená)

- G – choroby nervového systému (G454 – Prechodná celková stráta pamäti)
- H – choroby oka a ucha (H571 – očná bolesť)
- I – choroby obehovej sústavy (H65.0 – akútny zápal stredného ucha)
- J – choroby dýchacej sústavy (J40 – zápal priedušiek)
- K – choroby tráviacej sústavy (K59.0 – zápcha)
- L – choroby kože a podkožného tkaniva
- M – choroby svalovej a kostrovej sústavy a spojového tkaniva
- N – choroby močovej a pohlavnej sústavy
- O – tehotenstvo, pôrod a šestonedelie
- P – choroby pri pôrode a po narodení
- Q – vrodené chyby
- R – subjektívne a objektívne príznaky nezatriedené inde (R51 – bolesť hlavy)
- S – poranenia a zlomeniny
- T – úrazy a komplikácie
- V – vonkajšie príčiny úmrtnosti
- W – vonkajšie príčiny náhodného poranenia
- X – iné poškodenia (X33 – obeť blesku)
- Y – udalosti s neurčeným úmyslom (Y96 – choroba s povolania)
- Z – faktory ovplyvňujúce zdravotný stav a styk so zdravotnými službami (Z02 - vyšetrenie na administratívne ciele).

Komplený zoznam diagnóz je v číselníku diagnóz[23].

1.4 Využitie informatiky

Dvojica lekárska správa a *MKN* nám umožňuje použiť kontrolované algoritmy strojového učenia *(Supervised Algorithms of Machine Learning)*. Kontrolované preto, lebo pred ich aplikáciou na neznáme správy je ich potrebné najprv natrénovať na správnych dvojiciach. Voľný štýl lekárskych správ zrejme predurčuje aj využitie spracovania prirodzeného jazyka *(Natural Language Processing – NLP)*. Hovoríme už o probléme klasifikácie textu.[3] [2]

Definície zmienených pojmov spolu s popisom procesu učenia sú vysvetlené v nasledujúcej kapitole.

2. Strojová klasifikácia textu

Lekárske správy klasifikujeme diagnózami, t. j. priraďujeme im diagnózy. Správy sú písané nejakou formou prirodzeného jazyka, teda vlastne klasifikujeme text. To všetko chceme robiť automaticky, preto k tomu využijeme stroj, v našom prípade počítačový program. Keď ešte pritom vezmeme do úvahy vlástnosti správ z 1.1.2, tak si uvedomíme, že aby mal taký stroj zmysel, musí sa vedieť učiť sám iba z toho, čo má k dispozícií.

Na úvod popíšeme strojové učenie. Ako sa dá chápať a krátko, na intuitívnej úrovni, popíšeme jeho najpopulárnejšie metódy učenia. Následne predstavíme problém klasifikácie ako celku a detailnejšie popíšeme všetky jeho časti.

2.1 Strojové učenie

Snahou strojového učenia je dať programu schopnosť učiť sa bez toho, aby bol pre danú úlohu explicitne naprogramovaný. Jedná sa o pomerne modernú oblasť umelej inteligencie riešiacu značné množstvo praktických problémov od samo učiaceho šachu cez rozpoznávanie tvári na fotkách až po automatického pilota. [3]

Jedno z najpopulárnejších delení algoritmov strojového učenia je deliť ich na tri hlavné oblasti. Kontrolované učenie *(Supervised learning)*, nekontrolované učenie *(Unsupervides learning)* a spätnoväzbové učenie *(Reinforcement learing)*.

Kontrolované učenie má dve fázy. Tréningovú a testovaciu. V tej prvej sa algoritmus trénuje na základe dvojíc vstup a správny výstup. V testovacej fáze potom program dostane vstup a sám ma poskytnúť správny výstup na základe predchádzajúcej fázy. Algoritmy kontrolovaného učenia nájdu svoje uplatnenie predovšetkým pri získavaní informácií *(Information Retrieval)*, rozpoznávaní vzorcov *(Pattern Recognition)*, rôznej klasifikácií ako aj v bioinformatike, rozpoznávaní reči, ručného písma, či spamu a podobne. Vďaka tomu, že správna odpoveď je väčšinou známa, môžeme algoritmy odmeniť alebo postrašiť, a tým typicky dosiahnuť lepšie výsledky. [3]

Nekontrolované učenie funguje na inom princípe. V tomto prípade otázky nie sú priame a nie je na nich ani správna či nesprávna odpoveď. Jeho myšlienku môžeme vstup algoritmu popísať aj takto: „Algoritmus, tu máš neoznačené dáta a skús v nich nájsť nejaké štruktúry, ktoré Ti prídu zaujímavé.“ Pravdepodobne najtypickejším zástupcom je *Clustering*, ktorý rozdeľuje dáta na akési triedy ekvivalencie na základe súvislosti, ktoré na prvý pohľad často ani nie sú viditeľné. Masovo sa s tým môžeme stretnúť pri dolovaní dát *(Data Mining)*. [3]. Medzi ďalšie známe príklad patrí *Adaptive Resonance Theory (ART)* a *Self–Organizing Map (SOM)* z neurónových sietí.

Ešte uveďme aspoň základnú myšlienku *spätnoväzbového* učenia, t. j. algoritmus sa učí na základe spätnej väzby z okolia alebo od lektora [3]. Spätnú väzbu z okolia si môžeme predstaviť, napríklad, keď si všimneme, že keď vojdeme do jamy tak spadneme. Alebo si spočítame, že električka príde za 8 minút a cesta na cieľovú zastávku jej trvá 2 minútu ale vieme, že pešo tam budeme za 5 minúť, tak sa rozhodneme ísť pešo, lebo tým ušetríme čas a pohyb nezaškodí. Ale ak sa nám zmení stav, napríklad sme veľmi vyčerpaní, naša chôdza je obmedzená alebo sa vláčime s ťažkým nákupom, tak sa rozhodneme počkať na električku.

Lektora si môžeme predstaviť ako trénera plávania, ktorí si všimne, že počas plávania kraula vytáčame celý trup, tak nám o tom povie a odporúča nám vytáčať iba hlavu, aby sme sa vytáčaním trupu zbytočne nebrzdili. V umelej inteligenciu sa tento prístup používa najmä v agentných systémoch.

2.2 Klasifikácia úvodom

Klasifikácia alebo tiež *kategorizácia* znamená priradenie triedy, tiež kategórie, t. j. priradenie prvku z nejakej množiny určitému objektu z danej domény. Objekty z danej domény budeme všeobecne nazývať inštancie. Pre lepšiu názornosť uveďme zopár príkladov.

Príklad 2.1. Predstavme si, že v sklade chceme roztriediť tričká podľa veľkosti XS, S, M, L, XL. V reči klasifikácie je danou doménou množina tričiek a veľkosti sú triedy klasifikácie.

Príklad 2.2. Chceme iba na základe textu určiť pohlavie autora. Doménu predstavujú texty a triedy sú dve. Žena a muž.

Príklad 2.3. Chceme na základe absolvovaných predmetov študenta určiť, na akú pracovnú pozíciu sa môže hodiť. Doménou sú absolvované predmety a pracovné pozície sú triedy klasifikácie.

Príklad 2.4. Máme lekársku správy a pýtame sa, či určujú alebo neurčuje diagnózu I10. Lekárske správy sú doménou a triedy áno a nie.)

Na základe príkladov sa dá vytušiť, že na problém klasifikácie by sa celkom mohlo hodiť použiť kontrolované strojové učenie. O priradení podľa danej triedy rozhoduje *klasifikačný algoritmus*. Ten, ako bolo spomenuté v podkapitole o strojovom učení má v takom prípade dve fázy. Tréningovú a testovaciu. Pre prvú zo spomínaných fáz sú vstupom dvojice, t. j. atribúty a trieda na ktorú sa bude klasifikátor trénovať. Formálnejšie, dvojica je objekt z danej domény a k nemu správne priradená trieda. Atribút je niečo, čo nejakým spôsobom charakterizuje objekt z domény voči vybranej triede.

V príklade 2.1 môžu byť atribútmi výška a šírka trička. 2.2 pomer prídavných a podstatných mien, frekvencia citosloviec, 2.3 priamo zoznam predmetov. V poslednom príklade zrejme rozumným atribútom bude to, či daná správa obsahuje slovo tlak, jeho skratu TK, spojenie vysoký tlak a číslo v tvare X/Y.

tričko	**výška**	**šírka**	**veľkosť**
1	100	40	L
2	40	30	XS
3	107	43	L
4	80	35	M
5	100	60	XL

Tabuľka 2.1: Príklad vstupu pre model pre príklad 2.1

Na základe týchto vstupov v tréningovej fáze, klasifikačný algoritmusv vybuduje model pre dané atribúty a triedu, resp. triedy. To znamená, že sa naučí ako veľmi záleži trieda na hodnotách atribútu.

2.3 Atribúty a ich výber

Atribúty *(Features, Attributes)* sú vlastnosti popisujúce daný objekt. V našom prípade je daný objekt lekárska správa. Za atribúty si môžeme zvoliť akúkoľvek jednotku z domény. V prípade textu aj počet znakov správy, samohlások, trojslabičných slov, súčet každého druhého znaku podľa poradia v abecede alebo každé štvrté slovo napísané od konca.

Pri výbere atribútov sa fantázií medze nekladú, avšak zrejme ani jeden z uvedených výberov atribútov v skutočnosti neurčuje diagnózu žiadnej správy (alebo to aspoň autor nevidí). Čím pred nami stojí otázka, ako určiť, ktoré atribúty sú vzhľadom k triede relevantné.

Pre dôležitosť výberu čo najrelevantnejších atribútov je ešte nutné podotknúť, že irelevantné atribúty majú na väčšinu algoritmov strojového učenia negatívny vplyv, a najrelevantnejšie atribúty sú vyberané človekom na základe hlbokého pochopenia problému a významu atribútov v ňom. [1].

Samozrejme, neuberá to na význame automatického výberu, či už kvôli rýchlosti alebo vzácnosti výskytu ľudí s hlbokým pochopením problému. Keďže správy sú písané prirodzeným jazykom, k výberu atribútov využijeme metódy z vedeckej oblasti *spracovávania prirodzeného jazyka (Natural Language Processing, skrátene NLP*).

2.3.1 NLP

Spracovávanie prirodzeného jazyka (NLP,Natural Language Processing) je oblasť umelej inteligencie zaoberajúca sa analyzovaním, pochopením a generovaním prirodzených jazykov, t. j. jazykov, ktoré ľudia medzi sebou používajú na komunikáciu, za účelom interakcie s nejakým strojom v písomnej a hovorenej forme. Skúma spôsoby interakcie medzi strojom a človekom za použitia ľudského jazyka. Zasahuje do oblastí ako umelá inteligencia, strojové učenie, teória formálnych jazykov, lingvistika, psycholingvistika, kognitívne vedy a filozofia jazyka. [2]

2.3.2 Modely atribútov

Model je (multi)množina atribútov popisujúca doménové prvky. Pripomíname, že doménové prvky sú v našom prípade lekárske správy.

Pravdepodobne jedným z najpoužívanejších modelov v NLP sú *n-gramy slov*, kde n reprezentuje počet slov idúcich bezprostredne po sebe. 1-gramy, 2-gramy a 3-gramy voláme tiež unigramy, bigramy a trigramy.

Bag of Words (BoW) je jednoduchý model, kde atribút je reprezentovaný unigramom spolu s jeho frekvenciou. Navzdory tomu, že tento model už z jeho povahy úplne ignoruje poradie slov a ich kontext, ako si môžeme všimnúť na príklade 2.5, pre jeho implementačnú jednoduchosť a dobré výsledky využíva sa pomerne často.

Príklad 2.5. *Text*: Eva má mamu. Mama má Evu. Eva má rada zmrzlinu. Mama nie. *Unigramy* sú Evá, má, mamu, Mama, Evu, rada, zmrzlinu. nie. *Bigramy* má mamu, má rada a podobne. Spolu z týchto atribútov dostávame *Bow*: {Eva : 3, má : 3, Mama : 2, mamu : 1, rada : 1, zmrzlinu : 1, nie : 1, má mamu:1, má rada:1 }

Bigramy oproti *BoW* pridáva jednoslovný kontext. Na prvý pohľad to nie je veľa, bigramy "nádor nájdenýä "nádor nenájdený"môžu mať pri určovaní diagnózy kľúčový priam význam.

Poznámka. *n-gramy* všeobecne možu byť zoskupenie n nejakých jednotiek po sebe. My však v celej tejto práci budeme n-gramy používať vo význame n-gramy slov.

2.3.3 Filter atribútov

V prechádzajúcej podkapitole sme predstavili modely, ktorých atribúty môžeme vyberať iba základe počtu ich výskytov. To nie je ideálny stav, lebo najpočetnejšie atribúty vôbec neberú do úvahy kategóriu (klasifikačnú triedu), v našom prípade diagnózu. Navyše, častokrát majú vysoký výskyt slová, ktoré sú potrebné pre gramatickú stavbu jazyka ale pre určenie kategórie majú veľmi nízku výpovednú hodnotu. Jedná sa prevažne o predložky, spojky či príslovky. Filtrovanie týchto slov je popísané v nasledujúcej kapitole a v tejto sa sústredíme na to, ako vybrať práve tie atribúty, ktoré vybranú triedu určujú. [4]

Nasledujuce metriky hovoria, ako veľmi ktoré atribúty predpovedajú kategóriu. Hovoríme, že merajú relevatnosť. Jednou z takých metrík je *Pointwise mutual information (PMI)* postavená na pomeroch klasickej pravdepodobnosti. Vyjadrené vzorcom:

$$PMI(a, K) = \log_2 \frac{P(a, K)}{P(a)P(K)} , \tag{2.1}$$

kde a je atribút, K kategória a P klasická pravdepodobnosť. Priamo z vzorca pre výpočet *PMI* vyplývajú určité neošetrené prípady. Konkrétne sa jedná o vysoké *PMI* atribútu, ktorý sa v texte vyskytuje iba raz. V prípade bigramu môže mať jediný výskyt nie len samotný bigram, ale je slová, z ktorých je zložený. [1]

Sofistikovanejší prístup pre určovanie relevantnosti atribútu vzhľadom ku klasifikačnej triede predstavuje *informačný zisk (Information Gain (IG)). Informačný zisk* je založený na *entropii. Entropia* v teórií informácií meria nečistotu (impurity) alebo neistotu (uncertainty) náhodného výberu z nejakej množiny. Pri rovnomernom rozdelení možnosti nejakého náhodného javu, teda každá možnosť má rovnakú pravdepodobnosť výberu, potom nečistota množiny (vo význame dominantného výskytu prvkov s konkrétnou hodnotou náhodného javu) a ako aj neistota nejakého konkrétneho náhodného výberu je maximálna. Minimálna je keď presne vieme, čo si pri náhodnom výbere z množiny vytiahneme.

Zapísane vzorcom

$$H(A) = -\sum_{i=1}^{n} P(A = a_i) \log_2 P(A = a_i), \tag{2.2}$$

kde A je náhodná veličina a $a_1, \ldots, a_n$ sú jej hodnoty.

Informačný zisk je rovný očakávanému zníženiu entropie po rozdelení množiny podľa daného atribútu. Inými slovami, keď je atribút v rovnakom zastúpení pri každej diagnóze, rozdelenie kolekcie podľa neho entropiu kolekcie nezníži. Opačne, výskyt atribútu iba pri jednej diagnóze entropiu zníži, lebo daný atribút je pre určovanie danej diagnózy relevantný. Tým pádom rozdelenie kolekcie podľa neho entropiu zníži.

Aby sme mohli informačný zisk definovať vzorcom, potrebujeme si ešte definovať *podmienenú entropiu*, t. j. náhodný výber javu za predpokladu, že už nejaký iný jav nastal. Tá je definovaná nasledovne

$$H(A|B) = \sum_{b \in B} P(B = b) H(A|B = b). \tag{2.3}$$

Potom *informačný zisk* je definovaný ako rozdiel entropie a entropie rozdelenej podľa náhodnej veličiny B. Teda

$$IG(A; B) = H(A) - H(A|B). \tag{2.4}$$

2.3.4 Predspracovanie textu domény

Dôvodom *predspracovania* alebo *normalizácie* doménového textu je skutočnosť, že náš stroj dokáže porovnávať slová iba znak po znaku. Ale lekárske správy sú písané v českom jazyku, teda slová sa skloňujú, časujú, privlastňujú a podobne. Jedno slovo ma typicky viac tvarov. Bohužiaľ, všetky tie tvary pre náš stroj sú úplne rôzne a nezávislé slová, čo prirodzene, nie je žiadúce, lebo rožnými tvarmi sa význam slová alebo tiež *sémantika* nemení. Preto sa ešte pred výberom atribútov z textu celý text normalizuje. Typickými technikami normalizácie sú

- prevod veľkých písmen abecedy na malé,
- odstránenie nealfanumerických znakov,
- lematizácia,
- odstránenie stopwords.

Náš stroj nepracuje zo znakmi ako takými ale pracuje s ich číselnými hodnotami. Preto prevod veľkých písmen na malé je priam nevyhnutný. Avšak ako každá, ani táto jednoduchá technika nie je stopercentná. Protipríklady sa väčšinou týkajú skratiek. Napríklad seminár Súťaž Talentovaných Riešiteľov Obľubujúcich Matematiku je známy pod skratkou STROM, čo je úplne iný význam ako strom v prírode.

Do odstránenia nealfanumerických znakov sa počíta predovšetkým odstránenie interpunkcie ale aj znakov ako '+', '-', ':', '@', '%', '*', zátvoriek a pod.

Lematizácia je postup prevodu jednotlivých slov na ich základný gramatický tvar. Jedným z takých postupov je *stemming*. *Stemming* je odstránovanie afixov, t. j. predpôn a prípon slov. Afixy môžu byť odstraňované na základe slovníka predpôn a prípon alebo na základe pravidiel pre konkrétny jazyk. Inou možnosťou na získane základného tvaru slova je použitie morfologického slovníka.

K predspracovaniu textu ešte patrí ignorancia alebo odstránenie tzv. *stopwords*. *Stopwords* sme prvý krát spomenuli v predchádzajúcej sekcii a je to zoznam slov, ktoré sa v texte vyskytujú často ale nijakým spôsobom neurčujú klasifikačnú triedu. Môžu byť dvoch typov. Buď vychádzajú priamo z jazyka ako takého a použije sa ich verejne dostupný zoznam. Pre český jazyk sú to prevažne predložky, zámena, príslovky, častice, citoslovcia, číslovky. Druhou možnosťou je nechať si vygenerovaťj stopwords priamo z tréningových dát. Pre popis rozdelenia

dát na trénigové a testovacie viď nasledujúcu podkapitolu a sekciu *Tréningové a testovacie dáta*.

Pre automatické generovanie *stopwords* sa používa algoritmus *IDF (Inverse Document Frequency)*. Je založený na myšlienke, že atribút, ktorý sa vyskytuje pri väčšom počte inštancií má pre klasifikáciu menšiu informačnú hodnotu ako slovo, ktoré sa vyskytuje pri menšom pošte ištancií. Napríklad pre našu doménu atribút subjektívní vs bubínek. Matematicky vyjadrené

$$IDF(w) = \log_{10} \frac{|S|}{|S_w|} , \tag{2.5}$$

kde w je slovo S sú všetky lekárske správy a S_w sú všetky správy obsahujúce slovo w.

2.4 Klasifikačný model

Klasifikačný model je niečo, čo rozhodne o klasifikácii do prislušnej triedy. pomeňme príklad s tričkami 2.1. Model si môže povedať, že keď je tričko modré, tak má veľkosť X. Alebo keď súčet širky a dĺžky dá viac ako 200 cm, tak to bude veľkosť S. Samozrejme, pri takom modeli nie je veľká šanca, že bude lepší ako náhoda.

Pri textovej klasifikácii sa používajú už existujúce a osvedčené všeobecné algoritmy ako rozhodovacie stromy (Decistion Trees), klasifikátory založené na pravidlách (Pattern (Rule)-based classifiers), SVM Classifiers, neurónové siete (Neural networks), klasifikátory založené na Bayesovom pravidle (Bayesians classifiers).[4]

Na presný rozbor jednotlivých algoritmov v tejto práci nie je priestor, preto popíšeme aspoň tie, ktoré sme v našich experimentoch skutočne použili.

2.4.1 Naive Bayes

Naive Bayes klasifikátor je založený na kombinácii Bayesovho pravidla a predpoklade o nezávislosti atribútov. Bayesovské pravidlo udáva súvislosť medzi podmienenou pravdepodobnosťou nejakého javu s opačnou podmienenou pravdepodobnosťou.

Majme klasickú pravdepodobnosť P, klasifikačnú triedu C a atribúty $a_1, \ldots, a_n$. Bayesovo pravidlo

$$P(C|a_1, \ldots, a_n) = \frac{P(C)P(a_1, \ldots, a_k|C)}{P(a_1, \ldots, a_n)}, \tag{2.6}$$

predpoklad o nezávislosti atribútov

$$P(a_i|C, a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n) = P(a_i|C), \tag{2.7}$$

čo spolu dáva vzorec pre výpočet pravdepodobnosti *Naive Bayes*

$$P(C|a_n, \ldots, a_n) = \frac{P(C)P(a_1, \ldots, a_k)|C}{P(a_1, \ldots, a_k)}. \tag{2.8}$$

Výsledkom je pravdepodobnosť triedy C pri použití atribútov $a_1, \ldots, a_k$. Ostáva určiť hranicu, kedy sa trieda priradí a kedy už nie. Štandardne sa pre každú

triedu, vypočíta *Maximum a posteriori estimation* z Bayesovkej štatisky a vyberie sa trieda s najvyššou hodnotou. [1]

Naive Bayes vyžaduje binárne atribúty, tzn. každý atribút môže nadobúdať iba dve hodnoty. Pri *Multinominal Naive Bayes* môže mať atribút viac hodnôt. V klasifikácii textu to môže byť výhodné pre dlhšie texty. [1]

Predpokladať nezávislosť atribútov v reálnych textoch je dosť naivné (odtiaľ aj názov Naive Bayes), preto sú veľmi dobré výsledky tohto algoritmu prekvapujúce. Vysvetlení prečo by tomu tak mohlo byť sa ponúka viacero. O jednom z nich je článok [19].

2.4.2 Rozhodovacie stromy

Algoritmy, ktoré si pri konštrukcii svojho modelu budujú rozhodovacie stromy nazývame jednoducho **rozhodovacie stromy**. Koreň a vnútorné uzly rozhodovacieho stromu reprezentujú atribúty. Listy predstavujú kategórie. Cesty od koreňa k listu zodpovedajú klasifikačným pravidlám.

Algoritmus konštrukcie rozhodovacieho stromu môžeme popísať nasledovne.

Nech $A = \{a_1, \ldots, a_n\}$ je množina atributúv, $K = \{k_1, \ldots, k_m\}$ je množina kategórií a S je tréningová sada dát z danej domény. Na začiatku vyberieme atribút $a_i \in A$ a umiestnime ho do koreňa. Potom vyrobíme vetvu pre každú z jeho hodnôt alebo intervalu hodnôť, čo S na podmnožiny podľa hodnôt atribútu a reprezentovaných hranami. Pre každú vetvu sa tento proces rekurzívne opakuje ale vyberá sa iba z atribútov, ktoré neboli použité ako uzol, t. j. $A \setminus \{a\}$. Koreň vetvy sa stáva listom ak:

- všetky jeho zodpovedajúce inštancie z tréningovej množiny S patria do jednej triedy,
- minuli sa všetky atribúty a jeho inštancie patria aspoň do dvoch rôznych tried. Potom sa koreň stane listom podľa jeho najpočetnejšej triedy,
- minuli sa prvky tréningovej množiny alebo žiadna inštancia nepatrí koreňu. Potom sa koreň stane listom podľa najpočetnejšej triedy svojho rodiča.

Otázkou ostáva ešte voľba atribútu. Prirodzene by sme chceli zvoliť taký atribút, ktorý rozdelí inštancie čo možno najrýchlejšie. Preto sa vezme atribút s najväčším informačným ziskom (IG). IG je zrejme najvyšší vtedy, ak všetky inštancie padnú do jednej triedy. Naopak, najnižší je keď inštancie padnú do tried rovnomerne. Nech V sú prvky množiny S na aktuálnej vetve, Z je množina podmnožín množiny V vytvorených po rozdelení atribútom a. Informačný zisk sa potom definuje ako

$$IG(a) = H(V) - \sum_{z \in Z} P(z)H(z), \tag{2.9}$$

kde H je entropia definovaná vzorcom 2.2.

Štandardtnou optimalizáciou je orezávanie rozhodovacieho stromu, tzv. *pruning*. Ak sa počas generovania stromu pre vybraný testovací atribút nepresiahne stanovený prah, rekurzia sa zastaví. To je pre-pruning. Post-prunik už na hotovom strome celú vetvu nahradí listom v prípade, že veľkosť očakávanej chyby sa po nahradení listom zmenší. [1]

Takto popísaný algoritmus konštrukcie rozhodovacieho stromu má názov *ID3*. Jeho nástupca je algoritmus *C 4.5*, ktorého open source implementáciu *J48* využívame v experimentoch.[1]. Rozdiely medzi *ID3* a *C 4.5* sú popísané v článku [6].

Ďalší algoritmus, ktorý sme taktiež použili v experimentoch je *Random Forest*. Jeho hlavnou myšlienkou je vybudovanie veľkého počtu rozhodovacích stromov a na každom jednom nechať klasifikovať danú inštanciu. Výsledkom sa určí hlasovaním. Každý strom hlasuje za nejakú triedu a výsledná trieda je tá s najväčším počtom hlasov. Viac informácií ohľadom algoritmu Random Forest je možné najsť napríklad v[16].

2.5 Evaluácia

Evaluácia je proces vyhodnocovania klasifikačného modelu. Jej metriky nám povedia, ako dobre model klasifikoval inštancie a či funguje lepšie ako náhoda.

Predtým, než definujeme základne metriky, popíšeme cenu za chybu a čo robiť v prípade, že je dát málo, vysvetlíme pojem Confusion matrix.

2.5.1 Confusion matrix

Pre pochopenie zvyšných pojmov tejto kapitoly je nutné najprv porozumieť tomu, čo predstavuje *Confusion matrix*. *Confusion matrix* je jeden zo spôsobov určenia výkonnosti algoritmu a jeho vizualizuje v prehľadnej tabuľkovej forme. Popíšeme tento pojem priamo na príklade priraďovania diagnóz lekárskym správam.

Nech klasifikačný algoritmus priraďuje diagnózu D. Diagnózu rôznu od D označme $nonD$. Pri priraďovaní diagnózy musí každá lekárska správa padnúť do jednej zo štyroch disjunktných množín. Množina *TP (True positive)* sú správy s diagnózou D, ktorým algoritmus priradil D. *TN (True negative)* sú správy s $nonD$, ktorým bola priradená $nonD$. Ostávajú ešte *FN (False Negative)* a *FP (False positive)*, teda správy s diagnózou D, ktorým bola priradená nonD a naopak.

nonD	**D**	**priradená**
|TN|	|FP|	nonD
|FN|	|TP|	D

Tabuľka 2.2: Schéma confusion matrix

Bohužiaľ, ako je pri maticiach zvykom, ani táto nemá svoje pevné usporiadanie a v rôznych literatúrach sú stĺpce a riadky rôzne poprehadzované.

2.5.2 Meranie úspechu klasifikácie

Celkový počet chybných klasifikácií vyjadruje *Error rate*.

$$Errorrate = \frac{|FP| + |FN|}{N}, \tag{2.10}$$

kde N je počet všetkých inštancií.

Precision vyjadrujeme presnosť klasifikácie nejakej triedy. Inými slovami, ak už je niečo klasifikané danou triedou, ako veľmi je to pravda. Recall alebo tiež citlivosť hovorí koľko veľa inštancií s danou triedou sme klasifikáciou podchytili.

$$Precision = \frac{|TP|}{|TP| + |FP|}, \tag{2.11}$$

$$Recall = \frac{|TP|}{|TP| + |FN|}. \tag{2.12}$$

O celkovom úspechu klasifikácie jedna veličina bez druhej nemôže hovoriť, pretože ak jedna rastie, tá druhá zvykne klesať a naopak. Pre co najvyšší Recall pre danú triedu nám stačí tou triedou klasifikovať všetky inštancie, čím pádom určite podchytíme všetky z nich. Pre najvyšši Precision nám stači klasifikovať iba jednu inštanciu. Ak bola inštancia klasifikovaná správne, máme najvyšší možný Precision.

Veličina dávajuca do pomeru Precision a Recall ako ich harmonický priemer sa nazýva *F-Measure*. Čím je *F-Measure* vyšší, tým menej Recall a Precision od seba utekajú.

$$F\text{-}Measure = \frac{2Precision.Recall}{Precision + Recall}. \tag{2.13}$$

V reálnom živote sa málokedy stáva, aby obe Precision a Recall boli rovnako dôležité. Predstavne si situáciu, že máme vyšetriť pacienta na nejakú chorobu. To, že vyšetríme niekoho, kto tú chorobu nemá, nám prekáža ďaleko menej, ako keby sme nevyšetrili nejakého pacienta, ktorý je skutočne ma. V tomto prípade viac preferujeme Recall. Predstavme si inú situáciu. Máme v obchode identifikovať zlodeja. Prirodzene, nikoho nechceme nahnevať ani uraziť, preto v tomto prípade ďaleko viac preferujeme Precision. Ak chceme zohľadniť túto preferenciu pri meraní uspešnosti klasifikácie, možeme použiť niečo ako vážený *F-Measure* označovaný ako F_β, kde β je reálne číslo. Vyjadrené matematicky

$$F_\beta = \frac{(\beta^2 + 1)Precision.Recall}{\beta^2 Precision + Recall}. \tag{2.14}$$

.

Ako môžeme vidieť vo vzorca, pri $\beta \in (0, 1)$ preferujeme Precision, a pri $\beta > 1$ zase Recall.

Veličiny boli citované zo zdroja [1].

2.5.3 Meranie náhody

Ďalšou zo základných metrík sa volá *Kappa*. Porovnáva výsledky modelu s výsledkami náhody pri rovnakom rozdelení inštancií do klasifikačných tried. 0 je absolútna zhoda, 1 je maximum. Označme počet správne klasifikovaných inštancii modelom P a počet správnej klasifikácie náhodou N. Potom

$$Kappa = \frac{P - N}{|S| - N}, \tag{2.15}$$

kde S je množina všetkých správ. Zo vzorca je vidieť, že pri absolútnej zhode je $Kappa$ rovná 0. [1]

ROC arena je ďalšia veličina porovnajúca z klasifikačný model s modelom fungujúcom na princípe náhody.

Krivka vyobrazuje výkonnosť klasifikátoru bez ohľadu na triedy a cenu chyby. Na osi y je znázornený *Recall* a os x zobrazuje $1 - Precision$.

Náhoda ma krivku $x = y$. Klasifikátor je tým lepší, čím viac je *ROC krivka* pritiahnutá k ľavému hornému rohu grafu a jeho presnosť sa meria pod *ROC* krivkou, čo si môžeme predstaviť nasledovne. Rozdeľme správy na pozitívne a negatívne voči diagnóze. Ak z oboch tried vyberieme náhodne po jednej správe, klasifikátor by mi mal na *area pod ROC* priradiť správne triedy. [1]

ROC je skratka pre *Receiver Operating Characteristic* z teórie signálov.[1]

2.5.4 Cena za chybu

Ako sme popísali v podkapitole 2.5.2 pri F_β, v reálnom svete je väčšinou preferovanejší *Recall* oproti *Precision* alebo naopak. Povedané inými slovami, chyby pri klasifikácií majú často rozličnú váhu. Presne to vyjadruje *Cost matrix*. Viď tabuľka 2.3.

non D	D	priradil
0	1	non D
5	0	D

Tabuľka 2.3: Príklad Cost Matrix

V reči diagnóz príklad hovorí, že cena za nepriradenie diagnózy správe, ktorá ju má je 5 krát vačšia ako priradenie diagnózy správe, ktorá ju nemá. Zrejme za správne priradenie nechceme platiť nič.

V experimentoch sme použili meta klasifikátor *MetaCost*,[2] ktorý ako vstup berie klasifikátor a *Cost matrix*, podľa ktorej pre klasifikátor upraví rozloženie dát.[5]

2.5.5 Tréningové a testovacie dáta

Sme vo fáze, kedy natrénovaný klasifikátor dáva dobré výsledky. Zaujíma nás, či je natrénovaný skutočne dobre alebo sme ho trénovali nejakou špecifickou sadou dát, tzv. *overfitting*. Pri *overfittingu* sa výsledky klasifikácie môžu značne líšiť.

Pre získanie odpovede potrebujeme klasifikátoru dať na vstup úplne inú sadu dát, tzv. testovacích dát. Samozrejme, testovacie dáta okrem domény nesmú mať s trénigovými dátami nič spoločné. Až na základe získaných výsledkov môžeme predikovať chovanie klasifikačného modelu na nových dátach.

V týchto súvislostiach sa ešte niekedy hovorí o *validačných* dátach. Tie sa používajú na ladenie už natrénovaného klasifikáčneho modelu. Samozrejme, v

[1] Počas 2. svetovej vojny radaroví operátori museli rozhodnúť či bod na radarovej snímke je nepriateľ, spojenec alebo iba šum. V teórií signálov sa táto schopnosť volá Receiver Operating Characteric. V roku 1970 sa teória signálov ukázala ako užitočná pri interpretovaní výsledkov medicínskych testov [20]

[2] MetaCost klasifikátor je súčasť WEKA.

prípade že výsledky klasifikácie na testovacích dátach sú odpovedajúce, môžu sa taktiež pridať do celkových tréningových dát.[1]

2.5.6 Ak je dát málo

Voľne dostupných dát v praxi býva málo. Organizácie si zvyknú svoje, rokmi nazbierané, dáta strážiť. Preto sa vymýšľali rôzne techniky, ako natrénovať klasifikátor, ak nie je k dispozícií dostatočné množstvo dát na vyrobenie tréningovej a testovacej sady.[1]

Jedna s najčastejšie používaných techník sa volá *n-fold cross-validation (CV)*. Na začiatku sa sada dát náhodne premieša a rozdelí na n sád rovnakej veľkosti, tzv. *foldov*. Potom sa v každej iterácií jeden fold použije na testovanie a zvyšok na trénovanie. To sa opakuje n - *1* krát aby každý fold bol použitý práve raz. Nakoniec sa výsledky spriemerujú. [1]

Typicky používa $n = 10$, pretože sa ukázalo, že 10 je to správne číslo na získanie najlepšieho odhadu chyby.[1]

Variácií N-fold *CV*, kde N je počet inštancií sa vraví *Leave-one-out*. Má tendenciu byť menej spoľahlivá, keďže testovacia inštancia nemusí mať s tréningovými nič spoločné.[1]

Bez bližšieho popisu spomeňme ešte metódu *bootstrap* umožňujúca tej istej inštancii byť v tréningových i testovacích dátach.

3. Experimenty

3.1 Dáta

Mali sme k dispozícií milión anonymizovaných dvojíc, [1] kde jedna dvojica sa skladá z lekárskej správy a jej priradenej diagnózy. Tieto dvojice sme náhodne rozdelili na tréningové a testovacie v rovnakom pomere.

Dáta pochádzajú s reálnych ambulancií a polikliník z celého územia Českej Republiky. Správy sú písané v českom jazyku.

tréning	**test**
522747	524943

Tabuľka 3.1: Počty náhodne rozdelených dát

3.1.1 Vlastnosti správ

Pre pohodlnosť čitateľa pripomeňme špecifické vlastností lekárskych správ z prvej kapitoly aj s príkladmi.

- informačný heslovitý štýl
- preklepy
- nedôsledné oddeľovanie slov
- skrátené slová a značky
- vlastné stopwords
- kontextová nadväznosť

Príklad 3.1. Subjektivní: Quick 1,64 INR - Warf 3mg 1-0-0 5x týdně, a 2 x týdně 1,5-0-0. Objektivní: KP komp. AS zdá se pravidelná.

Príklad 3.2. Závěr: Ukončena PN k datu 23.11.2010.

Príklad 3.3. Subjektivní: Cítí sedobře

Príklad 3.4. Subjektivní: Průjem jako její přítel. Špatně spí. Objektivní: KP komp. Závěr: Kontrola p.p., při průjmu. Dieta nutná. Poučena o rizicích Hypnogenu. Neukázněný pacient.

Príklad 3.5. Objektivní: 6 250gr.Srdce,plice bpn.Vidí,slyší.Pevně drží hlavičku.Nutrilon prem.

[1] mená autorov lekárskych správ a pacientov sú anonymné, ale pri takom množstve dát nebolo v našich silách anonymizovať mená lekárov priamo v texte správ. Manuálnym prejdením stoviek správ usudzujeme, že ich výskyt je zriedkavý.

Príklad 3.6. Objektivní: Krvácí z nosu opakovaně,od včera silné bolesti.Hypertonička, kardiačka,Diabetička,Asthma bronchiale.Bere Tenaxum,Ramil,Moduretic,Verogalid. Alergie na PNC,ACP.RTG PND:jen lehké závoj.zastření zevní l/2 pravé front.dut.,bez zn.hladinky tekutiny.6 v.s.6 6V6-W-+R++Cl-C5+Dg:sinusitis front.cat.l.dxxDop:nosní kapky,promazávat nos O-Framykoinem,Rovamycine tabl3MUI 2xl,kontrola za týden,při nez.stavu za 5 dní.Vincentku

Príklad 3.7. Objektivní: Bolesti v krku ustoupily.Obj:sliznice laryngu a tracheiiklidnější.Tonsilly palat.s oj.chron.čepy.Jodglycerin lok.Dop:jodglycerin kloktat,ko v červnu.

Príklad 3.8. Subjektivní: Kontrola, teplolots nejsou zlpešení . Objektivní: Hrdlo klidnél plíce čisték ostatní nrma Závěr: Zlpešení doužívat léky, kontrola dle potřeby nyní verbebrogenní problémy

Informačný heslovitý štýl je dobre pozorovateľný na všetkých príkladoch. Preklepy v slovách vidíme najmä na poslednom príklade. Odhadom však môžeme povedať, že väčšina správ je bez zjavných preklepov. Nahradzovanie štandardných oddeľovačov slov, bodkou, čiarkou alebo dvojbodkou pri oddeľovaný slov je pomerne časté, viď príklady 3.5, 3.6, 3.7. V ťažko odhadnuteľnej miere sa vyskytujú aj slová písané bez oddeľovača pospolu. Napríklad *tracheiiklidnější* z príkladu 3.7

Z rozhovoru s lekárom vyplynula neexistencia všeobecných značiek, ktoré by sa učili na lekárskych školách. Avšak medzi lekármi existujú pomerne zaužívané značenia ako napr. V *6/6* (vidí so vzdialenosti 6 metrov), *C++* (cukor pozitívny), *R++* (výbavné reflexy), *W* (nadváha), *S* (sirota). *C1 až C7* zvyknú ortopédi alebo chirurgovia označovať krčne stavce, *Th1 až Th12* hrudné stavce atď. Existujú aj zdravotnícke zariadenia s vlastným verejným zoznamom používaných značiek používaných v ich lekárskych dokumentáciách. Lekári s obľubou používajú aj skrátené slová bežného charakteru, napríklad *Klíšť. encefaltída, cibul. štáva, kontrola v Ut* a podobne.

Okrem *stopwords* pre český jazyk dáta obsahujú desiatky slov typické pre doménu lekárskych správ. Typickými príkladmi sú slová *subjektívní*, *objektívní*, *závěr*, *nález*, *odběr*, *predskripce*, *norma*, *manželka* atď.

Lekár musí napísať správu pri každej pacientovej návšteve, čím vzniká veľké množstvo nových správ, ktoré sa ale vzťahujú na tie predchádzajúce a samotné nové správy ako také nemajú dostačujúci obsah pre určenie akejkoľvek diagnózy. Typickými príkladmi sú 3.2 a 3.3. Podobných správ je odhadom mnoho.

3.1.2 Najčastejšie diagnózy

V tabuľke 3.2 sú najčastejšie diagnózy z tréningových a testovacích dát. V 3.3 sú pomery diagnóz z tréningových dát o sto tisíc záznamov, ktoré boli použité pre trénovanie klasifikačných modelov z dôvodu výpočtovej náročnosti.

kód	**popis**	**tréning %**	**test %**
I10	Esenciální (primární) hypertenze	10,5	10,7
Z000	Celkové lékařské vyšetření (prohlídka)	4,9	4,9
Z001	Rutinní zdravotní prohlídka dítěte	4,0	4,0
J069	Akutní infekce horních dýchacích cest NS	2,7	2,7
J00	Akutní zánět nosohltanu	2,6	2,6
J039	Akutní tonzilitida NS	2,6	2,6
J209	Akutní bronchitida NS	2,5	2,5
J029	Akutní zánět hltanu NS	2,4	2,4
I259	Chronická ischemická choroba srdeční NS	1,9	1,9
R69	Neznámé a neurčené příčiny nemocnosti	1,5	1,5
K30	Funkční dyspepsie	1,1	1,1
Z008	Jiná celková vyšetření (prohlídky)	1,0	1,0
Z029	Vyšetření pro administrativní účely NS	1,0	1,0
I48	Paroxyzmální fibrilace síní	1,0	1,0
H660	Akutní hnisavý zánět středního ucha	0,9	0,9
E119	Diabetes mellitus nezávislý na inzulinu bez komplikací	0,8	0,8
Z235	Potřeba imunizace proti samotnému tetanu	0,8	0,8
J111	Chřipka s jinými projevy na dýchacím ústrojí, virus neidentifikován	0,7	0,7
J040	Akutní zánět hrtanu	0,7	0,7
J010	Akutní zánět čelistní dutiny	0,7	0,7

Tabuľka 3.2: Počty najčastejších diagnóz

P.	kód	popis	%	H%
1.	I10	Esenciální (primární) hypertenze	8,9	8,9
2.	Z001	Rutinní zdravotní prohlídka dítěte	4,2	13,1
3.	H660	Akutní hnisavý zánět středního ucha	4,0	17,1
4.	J00	Akutní zánět nosohltanu	3,9	21,0
5.	H681	Obstrukce Eustachovy trubice	2,9	23,9
6.	J010	Akutní zánět čelistní dutiny	2,9	26,8
7.	J209	Akutní bronchitida NS	2,8	29,6
8.	H903	Percepční nedoslýchavost, ztráta sluchu oboustrnná	2,4	32,0
9.	J069	Akutní infekce horních dýchacích cest NS	2,4	34,4
10.	J039	Akutní tonzilitida NS	2,4	36,8
11.	Z000	Celkové lékařské vyšetření (prohlídka)	2,3	39,1
12.	J029	Akutní zánět hltanu NS	1,8	40,9
13.	J040	Akutní zánět hrtanu	1,7	42,6
14.	I259	Chronická ischemická choroba srdeční NS	1,2	43,8
15.	H650	Akutní serózní zánět středního ucha	1,2	45,0
16.	K30	Funkční dyspepsie	1,1	46,1
17.	I159	Sekundární hypertenze NS	1,0	47,1
18.	E119	Diabetes mellitus nezávislý na inzulinu bez komplikací	1,0	48,1
19.	R040	Krvácení z nosu - epistaxis	0,9	49,0
20.	I48	Paroxyzmální fibrilace síní	0,8	49,8

Tabuľka 3.3: Počty najčastejších diagnóz v použitej tréningovej sade o 100 000 lekárskych správach so stĺpcami kód, popis, poradie, percentil výskytu, hromadný index

3.1.3 Pôvod dát

Výhradne na účely tejto práce mi dáta poskytla česká firma TERSINIDA CZ a. s. a pre verejnú demonštráciu tejto práce zverejnila tisíc lekárskych nálezov s priradenými diagnózami, ktoré sú k dispozícií na priloženom CD.

3.2 Vybrané diagnózy

Český číselník diagnóz [23] obsahuje viac ako desať tisíc diagnóz. Z toho dvadsať diagnóz vymenovaných v tabuľke 3.2 je priradených takmer polovici správ, ktoré sme mali k dispozícií.

Najčastejšie sa vyskytujúca diagnóza, okolo 10%, je diagnóza *I10*. Druhú a tretiu priečku obsadili preventívne prehliadky *Z000* okolo 5% a *Z001* okolo 4%.

Budovanie klasifikačného modelu na veľkom počte záznamov je výpočtovo, náročné[2] preto sme vybrali päť diagnóz, pre ktoré sme hľadali ten najlepší model. Pri ich voľbe sme vychádzali so spomínaných tabuliek 3.2 a 3.3 a zvolili sme tieto:

[2]pri výkone dnešného bežného PC sú to rádovo dni

- I10 – Esenciální (primární) hypertenze,
- Z001 – Rutinní zdravotní prohlídka dítěte,
- J00 – Akutní zánět nosohltanu,
- H660 – Akutní hnisavý zánět středního ucha,
- K30 – Funkční dyspepsie.

Popíšeme práve vybrané diagnózy, vyriekneme očakávané výsledky a ich dôvody ako aj dôvody samotného výberu.

3.2.1 I10 – Esenciální (primární) hypertenze

Signifikantne najfrekventovanejšiu diagnózu *I10*, ľudovo tiež nazývanú *vysoký tlak*, nemôžeme vynechať. Z bežnej laickej skúsenosti by sa pravdepodobne dali očakávať atribúty ako *tlak*, *vysoký*,*TK*, *hypertenze*, *meřit*, *KP komp.* , *doma* a predovšetkým výsledok merania tlaku. Meranie tlaku je vo formáte *TK systola/diastola*. Bohužiaľ, ako s každým iným aj s posledným vymenovaným atribútom pracujeme ako s textom, t. j. nemôžeme diagnózu strojovo určiť jednoduchým pravidlom viac ako 120/80.

V tréningových správach s diagnózou *I10* sa často vyskytujú spojenia ako *TK systola/diastola*, *KP komp.*, *Bez akut. stesků*, *Bez potíži*, *Kontrola*, *Dnes*, *medikace*. Predpokladáme, že práve táto diagnóza ma najväčší počet správ s chýbajúcim kontextom. V každom prípade, kvôli podielu celkových správ očakávame dobrý *Recall*.

Príklad 3.9. Subjektivní: Bez akut. stesků. Objektivní: TK 160/95, KP komp. Terapie: Prestance 10/5 1-0-0, Flexove

Príklad 3.10. Subjektivní: Bez akutních stesků. Chodí na gastroenterologii - hemoroidy. Objektivní: KP komp. TK 130/80, měří si doma 120/65, obesitas.

Príklad 3.11. Subjektivní: Cítí se dobře Objektivní: TK 130/80

Príklad 3.12. Subjektivní: Přichází pro léky. Závěr: Kontrola p.p.

Príklad 3.13. Subjektivní: Měla virosu. Občas jí bolí za krkem. Závěr: Zkusí vyměnit polštář.

Poznámka. Pri klasifikácií textu berieme každú hodnotu ako reťazec znakov. Preto aj meranie tlaku je iba reťazec znakov, hoci by bolo výhodnejšie pracovať s hodnotami nízky, normálny, vysoký.)

3.2.2 Z001 – Rutinní zdravotní prohlídka dítěte

Detskú zdravotnú prehliadku, sme vybrali pre jej predpísanú štruktúru a dobré zastúpenie. Bez filtrovania atribútom, v našom prípade *PMI* alebo *IG* neočakávame výsledky na úrovni iných diagnóz.

Nepredpokladáme, že najpočetnejšie atribúty sú napríklad *držení tela*, *kostra*, *genitál*, *akce srdeční*, *končetiny normální*, *končetiny*, *pubes*. Naopak, za pri výbere

atribútov pomocou filtra predpokladáme lepšie výsledky oproti ostatným diagnózam. Pri typickej dĺžke správ môžeme predpokladať, že každé navýšenie počtu atribútov bude pozitívny dopad na výsledky.

Príklad 3.14. Subjektivní: Bez potíží Objektivní: Lucidní, orientována místem a časem. Kostra gracilní, svalstvo slabší. Kůže, aniktericka, bez patol. eflorescencí, podkožní tuková vrstva slabá. Ochlupení přítomno jemné. Výživa dobrá. Držení těla dobré. Hlava mezocephalická, oči s bulby paralelními, skléry bílé, zornice izokorické, spojivky růžové bez sekrece. Nos bez viditelné sekrece, volně průchodný. Uši s boltci normálně konfigurovanými, zvukovody bez sekrece. Ústa souměrná, chrup zdravý, jazyk plazí středem, hrdlo lehce prosáklé, tonsily v obloucích, bez obsahu. Krk souměrný s neomezenou hybností, lymfatické uzliny nehmatné, štítná žláza nehmatná. Hrudník souměrný, dýchá v celém rozsahu. Mamily v úrovni hrudníku. Akce srdeční pravidelná, ozvy bez vedlejších šelestů, dýchání sklípkové. břicho v úrovni hrudníku, dýchá v celém rozsahu, měkké, prohmatné, bez patologických rezistencí, bez peritonismu, hepar v oblouku, lien nehmatný. Tříselné krajiny bez patologie. Genitál dívčí bez výtoku. Pubes nepřítomné. Končetiny normální konfigulace, volně pohyblivé ve všech kloubech. SKN. Moč B0 C0.

Príklad 3.15. Subjektivní: Bude se hlásit na gymnázium, v péči alergologie a imunologie pro snížení Th Ly a zvýšení IgE a polinosu Objektivní: Lucidní, orientován místem a časem. Kostra dobře vyvinuta, svalstvo dobře vyvinuté. Kůže, aniktericka, bez patol. eflorescencí, na zádek kroužkovité eflorescence. jizvy po varicelle,podkožní tuková vrstva silnější. Ochlupení přítomno jemné. Výživa dobrá. Držení těla dobré. Hlava mezocephalická, oči s bulby paralelními, skléry bílé, zornice izokorické, spojivky růžové bez sekrece. Nos bez viditelné sekrece, volně průchodný. Uši s boltci normálně konfigurovanými, zvukovody bez sekrece. Ústa souměrná, chrup sanován, jazyk plazí středem, hrdlo lehce prosáklé, tonsily v obloucích, bez obsahu. Krk souměrný s neomezenou hybností, lymfatické uzliny nehmatné, štítná žláza nehmatná. Hrudník souměrný, dýchá v celém rozsahu. Mamily v úrovni hrudníku. Akce srdeční pravidelná, ozvy bez vedlejších šelestů, 1. lehce protažená, dýchání sklípkové. břicho v úrovni hrudníku, dýchá v celém rozsahu, měkké, prohmatné, bez patologických rezistencí, bez peritonismu, hepar v oblouku, lien nehmatný. Tříselné krajiny bez patologie. Genitál d chlapecký, varlata sestouplá, předkožka volná. Pubes nepřítomné. Končetiny normální konfigulace, volně pohyblivé ve všech kloubech. SKN. Moč B0 C0. lehce asymetrické držení těla, susp. zkratek LDK

3.2.3 J00 – Akutní zánět nosohltanu

Diagnóz z kapitoly dýchacej sústavy (začínajúcich písmenom J) je najviac a sú pomerne časté. Bez nejakého hlbšieho dôvodu sme vybrali *akutní zánět nosohltanu*. Relevantnosť lekárskych správ vzhľadom ku kontextu pozorujeme oproti *I10* v oveľa väčšom počte. Z nášho laického pohľadu je nám väčšinou so správ jasné, že diagnóza začína písmenom *J*, avšak osobne v texte nevidíme výrazný rozdiel v konkrétnych diagnózach.

Príklad 3.16. Nachlaz. a rýma

Príklad 3.17. Subjektivní: Krk a nos bez patogen.

Príklad 3.18. Objektivní: Vystavena PN od 14.12. ko 22.12. kolem 9 h.,zvracel,poslán do lab.,krev,moč a výtěry krk a nos,CRP pod 8.Ležet ev. Paralen-před týdnem nachlazen a horečky.

Príklad 3.19. Subjektivní: Odběr sputa na bakteriologii

Príklad 3.20. Subjektivní: Dnes byl na kontrolním vyšetřní na chir. amb, asi zůstal na rehab bude chodit na VK nyní bolí v krku, kašel, rýma neměřil teplotu Objektivní: T v ord hrdlo červené, ausk. norm zakoupí Paralen kapky na kašel

3.2.4 H660 – Akutní hnisavý zánět středního ucha

Diagnóza akútneho hnisavého zánetu stredného ucha má výrazný rozdiel v podiele výskytu v tréningovej a testovacej sade. V tréningovej sade dát je pomerne vysoká (4%), ale celkovo má 0,9%. Predpokladáme, že testovacie výsledky by sa nemali výrazne odlišovať.

Pri tejto diagnóze sme z dát vypozorovali častý výskyt atribútu *6V6-W-+R++Cl-C5*, ktorý sa po odstránení nealfanumerických znakov rozpadne na viacero atribútov ale nečakáme výrazne zlepšenie výsledkov, lebo atribúty budú stále spolu.

Príklad 3.21. Objektivní: Bolesti lev.ucha,svědění.Obj:vpravo středouší zavlhlé,vlevohlenohnis.sekrece.Toaleta.Sliznice HCD s chronlzá změnamiDop:čistit ucho 3% H202,Pamykoin gttae 3xd do ucha a donosu,Doxyhexal tabl 2OO l tabl denně.Ko po vybr

Príklad 3.22. Objektivní: Bubínek vlevo se diferencuje,vpravo klidný.6 v.s.6 6V6-W-+R++Cl-C5+ Dop:apky dokapat,ko p.p.

3.2.5 K30 – Funkční dyspepsie

Funkčnú dyspesiu alebo ľudovo poruchy trávenia sme vybrali ako jediného zástupcu kapitoly chorôb *K – choroby tráviacej sústavy* z tabuľky 3.2 a kvôli rovnakému pomeru v tréningovej sade dát ako aj celkovo. Ako atribúty sa dajú očakávať orgány tráviacej sústavy a ich produkty. Zo skúmania dát sme tiež nadobudli pocit značného počtu sprav bez klasifikačného kontextu.

Príklad 3.23. Subjektivní: Bolesti břicha přijdou a odezní - trvá to cca spontánní potrat 13.11. ve 4.týdnu gravidity. Byla na gynekologii ve FTN. A také bolesti zad uprostřed cca od listopadu. Spojuje si to. Apetit +, ale přiznává, že občas na zvracení. Teplotu neměřila. Minulý týden zimnice. Průjem 0, zácpa 0. Stolice normální. Pokašlává. Močení v pořádku, nepálí, nebolí, noc. Objektivní: Břicho aperitoneální, hepar 0, lien 0. Palpačně citlivost nad symfýzou. Tapottement negativní. Palpačně citlivé trny TH-L přechodu. Měla hodně stažené kalhoty v pase. Verucca na levém boku. CRP nižší než 8 mg/l. Závěr: Kontrola na gynekologii. Vystavena PN. Klid na lůžku. CRP nižší než 8 mg/l. Kontrola u nás 13.12.2010. Ad labor KO + diff, CRP zavolá si. CRP nižší než 8 mg/l.

Príklad 3.24. Subjektivní: TK 160/80...155/80, doma TK ale v normě

Príklad 3.25. Subjektivní: Bez akut stesků. Stále průjem. Objednal se na kolonoskopii zač prosince. Držel bezlepkovou dietu, snad prý trochu zlepšení. Objektivní: KP komp. Závěr: Kultivace stolice příště. Žádanka vydána. Ukončena PN, kterou měl od 2009. Je nezaměstnaný.

Príklad 3.26. Objektivní: TK 140/80 přichází pro léky, na IV prstu PDK v.s. kuří oko s výrůstkem -ad pedikůra či kožní

Príklad 3.27. Subjektivní: Od dnes rána mala prujem,asi 3x, 1x vracala, dietnu chybu neudáva, včera večer mala zimnicu Objektivní: Hydratovaná dobre, brucho priehm, hepar lien v obluku, bez palp.bolestivosti, jazyk povl ečený, nosohltan kludný,šija volná, bez mening. príznakov. TR Terapie: Probiotika, tekutiny, dieta, kontrola s výsledkom, mimo kolektiv(chodí na prax do školskej jedálne ako kuchárka)

Príklad 3.28. Objektivní: FOB pozit.

3.3 Použitá klasifikácia

Výsledky klasifikácie závisia na

- použitých dátach s zvolených tried klasifikácie,
- charakteristike správ,
- predspracovaní dát,
- výbere atribútov,
- voľbe kvalifikátoru.

Dáta a výber diagnózy sme rozobrali v prechádzajúcej podkapitole. V tejto podkapitole podrobnejšie popíšeme zvyšné body zoznamu.

Pre jednoduchosť odkazovania sa v texte zavedene všeobecný pojem technika pre každú jednu možnosť položiek zo zoznamu. Napr. PMI, stopwords, morfológia atď.

3.3.1 Charekteristika správ

Lekárske správy charakterizujeme binárnym vektorom vybraných atribútov. To znamená, že hodnoty vektoru signalizujú, či správa daný atribút obsahuje alebo neobsahuje.

3.3.2 Predspracovanie dát

V predspracovaní správ sme vyskúšali každú techniku aby sme zistili, aký veľký má tá ktorá technika dopad. Teória je popísaná v sekcii 2.3.4.

Prvou možnosťou je atribúty skladať iba slov oddelených bielymi znakmi. To značí zachovať interpunkciu a iné nealfanumerické znaky. Takto sa zachovajú napríklad vyššie spomínane slovo *6V6-W-+R++Cl-C5*, dátumy a desatinné čísla aj

aj slovo *lev.ucha,svědění* z príkladu 3.21. Doplnkom k tejto technike je odstránenie nealfanumerických znakov. Týmto napríklad slov *Paralen-před* v príklade 3.18 rozdelíme na slová *paralen* a *před*.

Čo sa týka lematizácie, otestovali sme vstavaný stemmerom pre český jazyk[8] a morfologický slovník[7].

Ako *stopwords* sme mohli použiť všeobecný *zoznam stopwords pre český jazyk*[9] alebo stopwords vygenerované priamo z dát pomocou *IDF*.

3.3.3 Výber atribútov

Typovo sme použili automaticky generované

- unigramy,
- bigramy,
- kombináciu unigramovv a bigramov.

Keďže *unigram* chýba akýkoľvek kontext, očakávame, že všeobecne najrelevantnejšie unigramy budú podstatné mena popisujúce symptómy, časti tela, lieky, zložky rôznych meraní alebo priamo časť názvu diagnózy. Z unigramovv ako *čistý*, *zelený*, *volný*, *užíva* o diagnóze nevieme povedať nič, avšak sila unigramovv spočíva v ich počte. Napríklad keď správa obsahuje unigramy ako *čisty*, *plíce*, *zelený*, *zápal*, *krk píchání* a *zvětšeny* pôjde pravdepodobne diagnózou súvisiacu s dýchacími cestami.

Bigramy už majú bezprostredný kontext. Ten sa dá v ideálnom prípade využiť na bližie skúmanie symptónov. Napríklad *tonsily zvětšeny* , *bílé povlaky* , *Warfarin 5mg*, *tk 190/100*, *vyšší diastola*, *hrdlo prosáklé*, *stavy únavy*.

Výber atribútov sme rozdelili do niekoľko typov.

- TOP – podľa frekvencie ich výskytu,
- PMI – podľa PMI2.1,
- IG – podľa IG2.4.

PMI a *IG* sme definovali v 2.3.3.

Poznámka. Budovať model klasifikačných algoritmov nad atribútmi, ktoré sa v celej sade dát vyskytli raz alebo dvakrát nie je veľmi smerodajné, preto sme stanovili minimálnu nastaviteľnú hranicu ich frekvencie. Rozumné minimum nám prišlo ako štyri. Týmto sme aj ošetrili vlastnosť *PMI*, kde atribút vyskytujúci sa iba raz má podľa vzorca na výpočet *PMI* 2.1 vysoké *PMI*.

3.3.4 Klasifikátory a cena za chybu

Pri výbere klasifikačných algoritmov v našom prípade zohralo úlohu niekoľko faktorov.

- možnosti programu *WEKA*[10],

- hrubá povaha dát,
- veľkosť dát,
- výkon bežného osobného počítača.

S uvážením všetkých vymenovaných faktorov, rozhodli sme sa v experimentoch používať rozhodovacie stromy *C 4.5* a *Random Forest* a *NaiveBayes*. [3] Tieto algoritmy sme púšťali sami o sebe ako aj s nastavenou cenou za chybu v prospech *Recall*. Viď *Recall*. Viď 2.5.4.

Poznámka. Cena za chybu bola pre rôzne algoritmy bola nastavovaná algoritmom *MetaCost*.

Poznámka. Všetky algoritmy v tréningovej fáze používali *10 - cross-validation*.

3.3.5 Evaluácia

Úspešnosť klasifikácie sme sa rozhodli merať cez *F-Measure*, presnejšie cez jeho variantu F_5.

Všeobecne sa nám javí prijateľnejšie priradiť diagnózu správe, ktorá ju nemá ako nepriradiť ju správe, ktorá ju skutočné má. Pripomeňme, že F_5 znamená päť násobne väčšiu prijateľnosť. Inými slovami, preferujeme *Recall* oproti *Precision*. To všetko za predpokladu, že *ROC* je výrazne vyššie ako 0,5.

Teoretické pozadie je popísane 2.5.

3.4 Fázy výpočtov

V tejto kapitole popíšeme, ako sa dostali k najlepším klasifikačným modelom a testovacie výpočty.

Úplne na začiatku sme dáta náhodne rozdelili na tréningové a testovacie. Až do poslednej fázy výpočtov, ktorá bola testovacia sme pracovali iba s tréningovými dátami.

3.4.1 0. fáza – Hľadanie klasifikátorov

Jednoduchým shell scriptom sme získali 500 najfrekventovanejších slov a chceli sme ich použiť ako atribúty. Na samotnú klasifikáciu sme chceli použiť externú kolekciu algoritmov strojového učenia známu *WEKA*[10]. WEKA ktorá pracuje so súbormi vo formáte ARFF. ARFF súbor obsahuje atribúty a správy charakterizované vektormi atribútov [11]. Tak sme si naprogramovali vlastný program, ktorý z dát vyberie najfrekventovanejšie slová a vyrobí z nich ARFF súbor. Časom sme vlastný program nazvali ARFFBuilder. Jeho dokumentácia je v prílohe.

Na získaný ARFF súbor sme pustili niekoľko klasifikátorov z WEKA a aplikovali ich pre diagnózy I10, I709 a J068. Najviac sa nám osvedčili Naive Bayes, J48 a Random Forest. Ostatné klasifikátory ako *Multilayer Perceptron*, *Bayes Net* alebo *Decision Table* mali výrazne horšie výsledky.

[3] implementácia algoritmu C 4.5 v programe WEKA má názov J48.

3.4.2 1. fáza – Hľadanie predspracovania a veľkosti atribútov

Náš program ARFFBuilder sme postupne rozširovali o odstránenie interpunkcie, pridali sme bigramy, český stemmer, až sme postupne pridali všetky techniky spomenuté v 3.3.2.

Každú z týchto techník sme aplikovali na dáta samostatne, aby sme získali odhad ich dopadu na výsledok. Nakoniec sme to všetko ešte zopakovali pre rôzne počty a typy atribútov a vybrané klasifikátory.

Nepredpokladáme, že by sa dopad rôznych techník predspracovania výrazne líšil medzi rôznymi diagnózami, ale dopady rôznych typov atribútov by sa už medzi diagnózami líšiť mohli. Preto sme to zopakovali ešte pre každú diagnózu zvlášť. Čo spolu dáva 1830 výsledkov.

Poznámka. Keď sme chceli posúdiť dopady všetkých kombinácii na výsledok, zvlášť pre každú diagnózu, potrebovali sme sa dívať na výsledky v ucelenej podobne. Preto sme si naprogramovali druhý program, ktorý pozbiera všetky výsledky za určitú diagnózu a dá ich do jedného CSV súboru.

3.4.3 2. fáza – Filtrovanie atribútov

Z prvej fázy výpočtov sme zistili, ktoré techniky vylučovacie techniky predspracovania (viď 3.3.2) môžeme skombinovať a očakávať aspoň rovnaké výsledky ako boli tie najlepšie v prvej fáze.

V žiadnej výpočtovej fáze sme zatiaľ nehovorili o filtrovaní atribútov, čo znamená, že sme pri ich výbere stále používali *TOP* filter. (viď sekciu 3.3.3). V tejto časti sme na tie najlepšie kombinácie tokenizácie a počtu atribútov použili okrem *TOP* aj *PMI* a *IG*. Pre ich vysvetlenie viď sekciu 3.3.3.

3.4.4 3. fáza – Testovacia

Na záver pustíme najlepšie klasifikačné modely z druhej výpočtovej fázy oproti testovacím dátam. Najlepší model chápeme z najlepším F_5 a ROC výrazne vyšším ako 0,5 podľa 3.3.5.

3.5 Výsledky

V tejto podkapitole popíšeme výsledky jednotlivých výpočtových fáz. Z časového hľadiska sme hlbšie analyzovali iba výsledky diagnózy I10. Druhým dôvodom je aj to, že I10 je vybraných diagnóz pre človeka bez zdravotníckeho vzdelania najpochopiteľnejšia a vidí v nej najviac súvislosti.

3.5.1 Značenie

Pokiaľ nie je povedané inak, pre celú túto kapitolu platí nasledujúce značenie. Obzvlášť pre grafy a tabuľky.

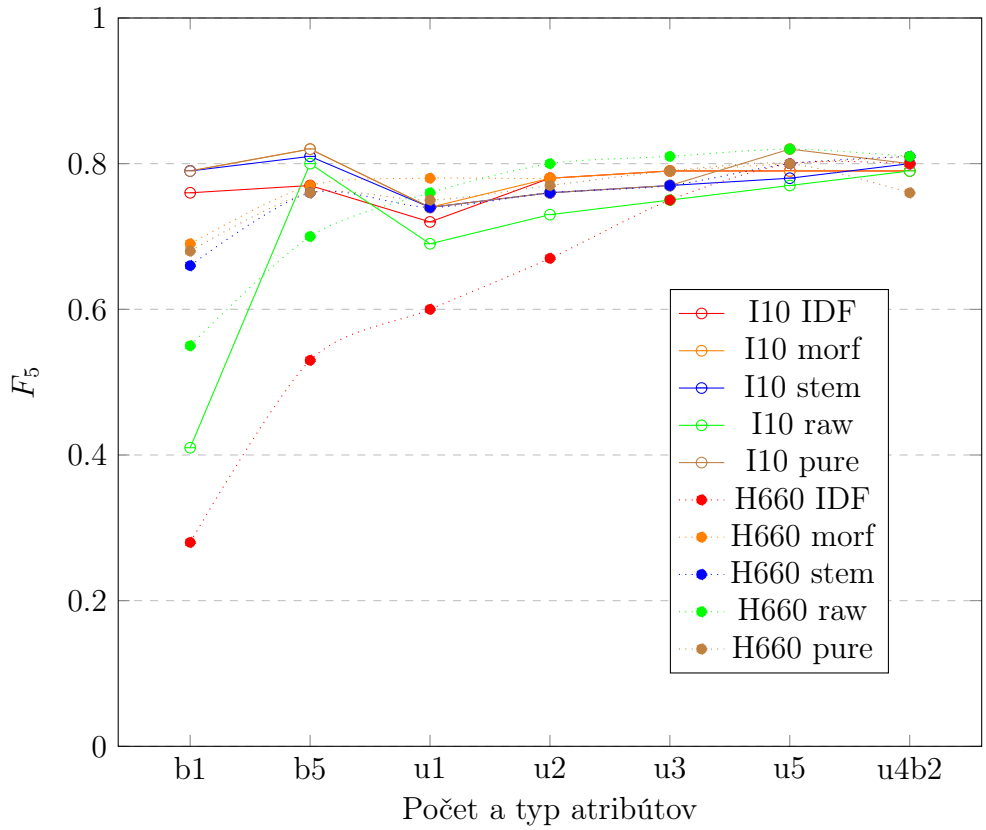
- uN = N00 unigramovv, kde je N je prirodzené číslo,

- bN = N00 bigramov, kde je N je prirodzené číslo,
- DG – diagnóza,
- stem – vstavaný český stemmer[8],
- morf – z dát vygenerovaný morfologický slovník[7],
- raw – zachovanie nealfanumerických znakov,
- pure – odstránenie nealfanumerických znakov,
- IDF – stopwords vygenerované na základe IDF,
- TOP – najfrekventovanejšie atribúty,
- IG – informačný zisk,
- PMI – vzájomná informácia,
- TN – True negative,
- FN – False negative,
- FP – False positive,
- TP – True positive,
- ROC – aréna pod ROC krivkou,
- K – Kappa,
- NB – Naive Bayes,
- J48 – implementácia C4.5 klasifikátoru,
- RF – Random Forest,
- MCRNB – Meta Cost s Naive Bayes a Cost Matrix FP = R a FT = 1,
- MCRJ48 – Meta Cost s J48 a Cost Matrix FP = R a FT = 1,
- MCRRF – Meta Cost s Random Forest a Cost Matrix FP = R a FT = 1,

kde N je prirodzené číslo a R reálne. Pre popis TN, FN, FP, TP viď sekciu 2.5.1 a pre popis Meta Cost a Cost matrix viď 2.5.4.

3.5.2 1. fáza

Ako sme spomínali v sekcii 3.4.2, výber predspracovania textu správ by mal mať dopad pre výsledky modelu. Nie je to úplne pravda, pretože drobné rozdiely vo vybraných diagnózach. Pre šetrenie priestorom uvedieme graf s najvyššími F_5 pre každú techniku a iba pre diagnózy I10 a H660. Všetky výsledky sú prístupné na priloženom CD.



Obr. 3.1: F_5 techník predspracovania textu pre diagnózu I10.

Z grafu vidíme, že pre sumárne výsledky na 100 000 predspracovanie textu nemusí mať nutne dopad. Ale pre náš je dôležité správne klasifikovať každú jednu správu, preto budeme ďalej používať odstránenie nealfanumerických znakov (v grafe ako pure) morfológiu (v grafe ako morf). Môžeme si do dovoliť, lebo ani jedna technika predspracovania F_5 podstatne neznižuje. Ale ako uvidíme neskôr, metrika F_5 je skôr pomocná. Čo sa týka ignorovania stopwords pri kóde diagnózy H660, vysvetlenie je jednoduché. Najfrekventovajšie bigramy a ani unigramy o počte menšom ako 300 bez stopwords generovaných IDF sa v správach kategorizovaných ako H600 nevyskytujú.

Top atribúty a stopwords podľa IDF

Najfrekventovanejšie unigramy sú klasické stopwords: *objektivní*, *v*, *subjektivní*, *na*, *bez*, *s*, *a*, *p*, *závěr*, *se*. Unigram *p* je najpravdepodobnejšie s lekárskeho termínu *kontrola p. p.*, v preklade *kontrola podle potřeby*. Prvý atribút ktorý môže mať význam pri určovaní diagnózy je *slezina* na 17. pozícií. Za ním na 30. priečke nasleduje *dýchaní*.

Najfrekventovanejšie bigramy sú: *v s*, *p p*, *oto negat*,*do nosu*, *v krku*, *kontrola za*, *objektívní hrdlo*, *v normě*, *dýchaní čisté*, *ko p*, *v noci*. Relevatné unigramy na popredných miestach v bigramoch sa umiestnili nasledovne: *nosu* 40. miesto, *krku* 36., *hrdlo* 23..

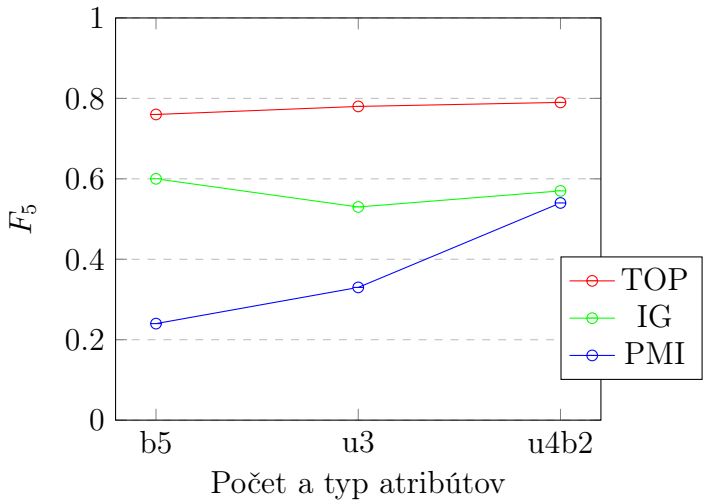
Stopwords, tým pádom že atribúty uvažujeme voči správe binárne, sú totožné s top unigramami. Obsahujú ďalšie medicínsky relevantné atribúty ako *tonsily*, *srdce*, *bubínek*, *nosohltan*, *laryng*, *pharyng*, *stolice*, *vincenta*. Generovali sme ich v počte 200.

3.5.3 2. fáza

Najprv sme z dát odstránili nealfanumerické znaky okrem znaku / kvôli merianiu tlaku a znaku - kvôli zápisu dávkovania liekov. Stopwords generované podľa IDF sme použili iba v prípade TOP filtru. Tieto stopwords obsahujú aj slová relevantné pre určenie diagnózy. Viď 3.5.2. Môže sa stať, že slová, ktoré sa môžu javiť ako bezvýznamové, môžu mať ako súčasť bigramu vysoké IG, a tak poukázať na nie úplne zjavné súvislosti.

V popisoch výsledku sme sa sústredili predovšetkým na IG, pretože PMI neberie do úvahy entropiu dát ani atribútu a navyše atribút s veľmi nízkou frekvenciou môže mať vysoké PMI.

I10



Obr. 3.2: Najlepšie filtrovanie atribútov k diagóze I10

Na grafe je jasne vidno, že TOP filter stabilne dáva najvyššie F_5. To je zvláštne, lebo TOP vôbec neberie do úvahy diagnózu. Vysvetlenie nám poskytne analýza vybraných atribútov a podrobná tabuľka výsledkov.

Na prvých priečkach výberu atribútov pomocou IG sú unigramy *tknout*, *ia*, slavná,[4] *preskribce*, *140/80*, *holter*, *150/80*, *150/90*, *130/70*, *130/80*, *160/90*, *145/80*, *stab*, *slavné*, *oxantil*.

[4]Skutočné meno lekárky sme anonymizovali za priezvisko Slavná.

tknout vyrobila morfologická analýza a pôvodné slová sú *tk*, *tK* a *Tk*, čo je skratka pre krvný tlak. *ia*, pôvodne *IA*, je skratka pre internú ambulanciu. *slavná* je priezvisko internej lekárky s frekvenciou 239, *Holter* je vyšetrenie krvného tlaku, *stab.* je skratka pre stabilizovaný krvný tlak a v správach sa skutočne vyskytuje predovšetkým v tejto súvislosti. *Oxantil* je liek na krvný tlak a merania tlaku sú zrejmé. Vysoký tlak má hodnotu nad 120/80.

Meno lekára napr. atribút *slavná* je dobrý prediktor, lebo lekár vďaka svojej špecializácii vyšetruje iba určité diagnózy. Ale je nutné poznamenať, že nemusí platiť pre inú dátovú sadu. Inak všetky atribúty sú všeobecne relevantné pri určovaní kódu diagnózy I10. Viď príklad 3.31.

Bigramy podľa IG: objektivní tknout, pro lék, tknout -, subjektivní pl, subjektivní pro, subjektivní lék, subjektivní preskribce, doktor slavná, objektivní pro, lék tknout, ia doktor, kontrola tknout, subjektivní tknout, tknout 140/80, elektrokardiograf sinus, subjektivní dom, subjektivní dobře, dom tknout, manželka pro.

Bigramy nám potvrdili slová jedného lekára, ktorého sme požiadali o binárne určenie I10 zo správ, a ktorý sa vyjadril, že ak správa hovorí iba o predpísaní liekov, tak má podozrenie na I10, lebo pri iných diagnózach sa do správy zvykne ešte niečo doplniť. Predložka *pro* naznačuje, že si pacient po niečo prišiel. Často práve pre lieky. *dom* bolo pre morfológiou . Pacienti si často merávajú tlak doma. Bigram *subjektivní dobře* napovedá o veľkom počte správ s chýbajúcou históriou pri I10.

Zistili sme, že kým unigramy mám IG viaže priamo na I10, bigramy dokážu poodhaliť súvislosti, ktoré nemusia byť na prvý pohlaď zrejmé. Čím sa nám zachovanie stopwords pre filtre atribútov, ktoré berú do úvahy samotnú kategóriu, ukázalo ako správne rozhodnutie.

To však stále úplne neodpovedá na otázku, prečo má TOP filter signifikantne lepšie F_5. Jedna pričíňa je vysoký podiel správ bez histórie ako ukazujú bigramy. A samozrejme, nie každá správa je historié je I10. Viď príklady 3.29 a 3.30. Ďalej vieme, že meranie tlaku patrí k bežných diagnostickým metódam v medicíne. V dôsledku toho pacient príde za lekárom s niečím iným a ako vedľajšiu diagnózu vysoký krvný tlak. Viď príklad 3.32.

Keď to je pravda, tak pre naše to znamená, že IG model musí mať výrazne vyššiu F_1. Podľa tabuľky podrobných výsledkov 3.4 je tomu skutočne tak. Najvyššie má b5IG-MCNB.

Príklad 3.29. I10 Nález: Pro léky

Príklad 3.30. N393 Nález: Pro léky

Príklad 3.31. I10 Objektivní: Odběr ad IA Dr. Slavná

Príklad 3.32. J209 Subjektivní: Kašel bolesti na prsou afebrilní v anamnese asthma bronchiále med. Symbicort Objektivní: Tk.: 160/100 p-68/min bronchit. poslech prodloužený výdech putridní spůtum Terapie: Klacid 500 2x1 Erdomed 2x1 Závěr: Ak, exacerbace chron bronchitidy

ID	TN	FP	FN	TP	P	R	F1	F5
b5TOP-MC5NB	23 225	67 863	135	8 811	0,12	0,99	0,21	0,77
b5IG-MC5NB	86 758	4 330	3 522	5 424	0,56	0,61	0,58	0,60
b5IG-NB	88 340	2 748	4 538	4 408	0,62	0,50	0,55	0,50
b5PMI-NB	90 437	651	6 817	2 129	0,77	0,24	0,36	0,25
b5PMI-MC5NB	90 629	459	7 155	1 791	0,80	0,20	0,32	0,21
b5TOP-NB	90 306	782	8 326	620	0,44	0,07	0,12	0,07
u3TOP-MC5NB	48 641	42 447	722	8 224	0,16	0,92	0,28	0,78
u3TOP-NB	64 407	26 681	1 923	7 023	0,21	0,79	0,33	0,71
u3IG-MC5NB	85 212	5 876	4 164	4 782	0,45	0,54	0,49	0,53
u3IG-NB	88 660	2 428	5 934	3 012	0,55	0,34	0,42	0,34
u3PMI-MC5NB	88 805	2 283	6 012	2 934	0,56	0,33	0,41	0,33
u3PMI-NB	90 173	915	7 459	1 487	0,62	0,17	0,26	0,17
u4b2TOP-NB	55 062	36 026	845	8 101	0,18	0,91	0,31	0,79
u4b2TOP-MC5NB	37 307	53 781	346	8 600	0,14	0,96	0,24	0,78
u4b2IG-MC5NB	86 799	4 289	3 878	5 068	0,54	0,57	0,55	0,57
u4b2IG-NB	87 120	3 968	3 924	5 022	0,56	0,56	0,56	0,56
u4b2PMI-MC5NB	85 348	5 740	4 115	4 831	0,46	0,54	0,50	0,54
u4b2PMI-NB	88 267	2 821	5 389	3 557	0,56	0,40	0,47	0,40

Tabuľka 3.4: Confusion matrix a metriky výkonu pre I10 v druhej výpočtovej fáze. **P** je Precision a **R** Recall.

Ďalej z tabuľky môžeme odpozorovať:

- TOP napriek tomu, že má všeobecne najlepší F5, neurčuje diagnózu. I10 priradí rádovo viac správam ako priradia PMI alebo IG, ale má najhoršiu presnosť.
- PMI je v prípade iba unigramov alebo iba bigramov presnejšie ako IG, ale nepodchytí až tak veľa prípadov, ako IG, ktoré je tiež pomerne presné.
- V prípade kombinácie unigramov a bigramov sú si PMI a IG blízke, ale predsa len IG má navrh.
- IG podchytí viac mierne viac ako polovicu správ s kódom diagnózov I10 a mierne viac ako v polovici prípadov ho priradí správne. Podľa atribútov vybraných IG sa dá predpokladať, že značne množstvo chybných klasifikácii má I10 ako vedlajšiu diagnózu a neklasifikované správy s I10 sú bez relevatného obsahu voči ľubovoľnej diagnóze.

Pre porovnanie vypíšme najlepšie atribúty PMI.

Unigramy: *levotyp*, *ifirmacombi*, *telmisartan*, *lusopr*, *irbesartan*, *lusopresu*, *polyartr*, *veloergometrie*, *krteat*, *amloratio*. Mnoho z nich sú lieky. Na 25. mieste sa umiestnil unigram *slavná* a pred ním na 16. mieste unigram *slavný*.

Bigramy: *- 125/70*, *lék potíž*, *- 150/90*, *- 135/70*, *ia objektivní*,*ebrantil 30*, *130/80 k*, *0-0-1 ob*, *neo 5*, *dnes 140/80*, *interna ichs*.

Aby bolo možné stratégiu výberu atribútov IG a PMI skutočne porovnať, je potrebné o nich vedieť dotadočné informácie ako ich frekvenciu, v akom počte inštancií s priradenou diagnózou sa vyskytujú a v koľkých správach celkovo. Pre vymenované údaje viď tabuľky 3.5 a 3.6.

Z nich je vidieť, ako PMI neberie do úvahy entropiu dát ale vyberá práve tie atribúty, ktoré určujú danú diagnózu čo najviac jednoznačne. Tieto atribúty sú svojim spôsobom jedinečné, lebo sa pri iných diagnózach takmer nevyskytujú. V našom prípade, hlavne kvôli nízkym frekvenciám výskytu, pôsobia, že majú platnosť skôr v rámci tréningových dát ako všeobecne. Pekným príkladom je bigram *lék potíž*. Pripomeňme ignoráciu atribútov s frekvenciou 5 a nižšiou. IG zohľadňuje entropiu dát, a preto atribúty s najvyšším IG majú rádovo vyšší počet výskytov, čím posilňujú ich všeobecný charakter.

Poznámka. PMI a IG v tabuľkách 3.5 a 3.6 sa líšia až na nižších radoch desatiných miest a boli zaokrúhlené na 4 desatiné miesta.

Atribút	**Freq**	**Dg**	**Správy**	**IG**
tknout	9435	4321	9435	0,26729
ia	413	325	413	0,26638
slavná	233	189	233	0,26618
preskribce	370	245	370	0,26615
140/80	685	373	685	0,26612
holter	255	146	255	0,26598
150/80	186	115	186	0,26598
150/90	261	145	261	0,26597
130/70	230	126	230	0,26595
130/80	1 120	506	1 120	0,26594
160/90	143	85	143	0,26594
145/80	99	64	99	0,26593
stab	97	63	97	0,26593
slavné	64	48	64	0,26593
oxantil	62	43	62	0,26592
objektivní tknout	3 672	2 154	3 672	0,26401
pro lék	1 737	1184	1737	0,26349
tknout -	343	292	343	0,26246
subjektivní pl	1 115	615	1 115	0,26239
subjektivní pro	464	324	464	0,26236
subjektivní lék	346	254	346	0,2623
subjektivní preskribce	362	245	362	0,26223
doktor slavná	211	172	211	0,26222
objektivní pro	264	183	264	0,26217
lék tknout	131	119	131	0,26216

Tabuľka 3.5: Atribúty s najvyšším IG pre diagnózu I10. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.

Atribút	**Freq**	**Dg**	**Správy**	**PMI**
levotyp	10	10	10	3,4831
ifirmacombi	8	8	8	3,4831
telmisartan	6	6	6	3,4831
lusopr	6	6	6	3,4831
irbesartan	5	5	5	3,4831
lusopresu	5	5	5	3,4830
polyartr	5	5	5	3,4831
veloergometrie	5	5	5	3,4831
krteat	5	5	5	3,4831
amloratio	15	14	15	3,38357
- 125/70	14	14	14	3,4831
lék potíž	13	13	13	3,4831
- 150/90	10	10	10	3,4831
- 135/70	10	10	10	3,4831
ia objektivní	10	10	10	3,4831
ebrantil 30	9	9	9	3,4831
130/80 k	9	9	9	3,4831
0-0-1 ob	9	9	9	3,4831
neo 5	9	9	9	3,4831
dnes 140/80	8	8	8	3,4831

Tabuľka 3.6: Atribúty s najvyšším PMI pre diagnózu I10. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.

H660

Uvedieme zoznam unigramov a bigramov s najvyššími IG a PMI pre diagnózu H660.

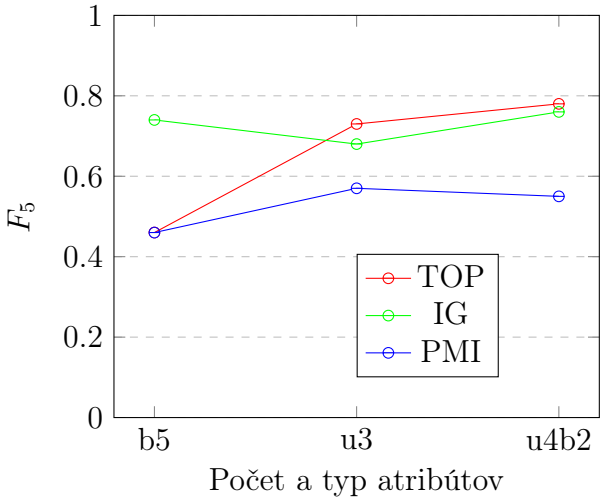
Atribút	**Freq**	**Dg**	**Správy**	**IG**
oma	845	716	845	0,08219
diferencovat	623	514	623	0,08172
pamykoin	1 541	793	1541	0,0814
paracentésa	281	274	281	0,08132
boralkohol	645	366	645	0,08107
bubínek	6 577	2 497	6 577	0,08103
lev	1344	615	1 344	0,08101
vyklenutý	266	194	266	0,08099
spont	254	189	254	0,08099
vyklenutí	419	248	419	0,08096
oma l	567	502	567	0,07805
vpravo bubínek	593	408	593	0,07759
s diferencovat	349	295	349	0,07753
pravit ucho	968	540	968	0,07753
lev ucho	935	516	935	0,07749
obj vpravo	978	521	978	0,07744
obj vlevo	758	424	758	0,07739
bolest pravit	449	302	449	0,07739
pamykoin gttae	1 073	539	1 073	0,07737
vlevo bubínek	474	305	474	0,07736
vlevo oto	843	443	843	0,07734

Tabuľka 3.7: Atribúty s najvyšším IG pre diagnózu H660. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.

Atribút	Freq	Dg	Správy	PMI
bubínekpřekrvený	21	21	21	4,635
sevyklenovat	17	17	17	4,635
perfdop	16	16	16	4,635
perfordop	16	16	16	4,635
shlenem	11	11	11	4,635
lehkézavlhnutí	9	9	9	4,635
provedenaparacentésa	9	9	9	4,635
ucho3	7	7	7	4,635
sediferencují	7	7	7	4,635
paracentésakrev	7	7	7	4,635
paracentésa krev	68	68	68	4,635
h202 pamykoin	41	41	41	4,635
kapka vyplachovat	29	29	29	4,635
paracentésa serosang	29	29	29	4,635
ucho augmentin	22	22	22	4,635
paracentésa bilat	22	22	22	4,635
diferencovat sluch	20	20	20	4,635
perforace serosang	18	18	18	4,635
překrvený začínat	18	18	18	4,635
začínat sevyklenovat	17	17	17	4,635

Tabuľka 3.8: Atribúty s najvyšším PMI pre diagnózu H660. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.

Z grafu 3.3 pozorujeme dominaciu TOP filtru, ale iba v unigramoch.



Obr. 3.3: Najlepšie filtrovanie atribútov k diagóze H660

J00

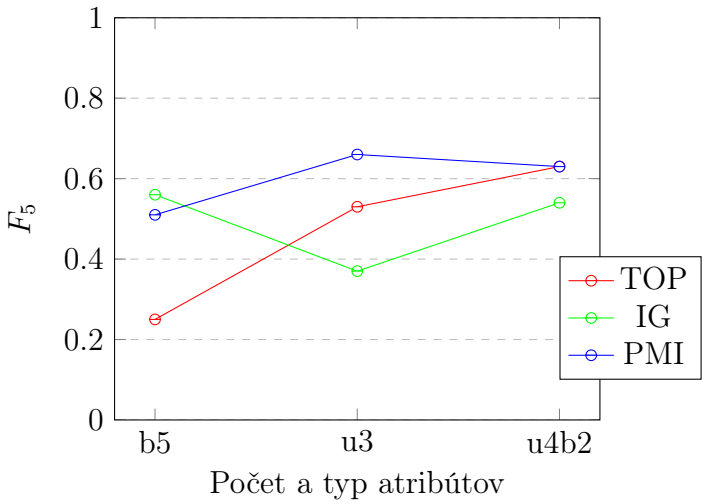
Uvedieme zoznam unigramov a bigramov s najvyššími IG a PMI pre diagnózu J00.

Atribút	**Freq**	**Dg**	**Správy**	**IG**
zostřený	742	494	742	0,07519
přenesený	455	276	454	0,07478
rudý	1533	682	1533	0,07477
hlenový	412	251	412	0,07474
rhinopharyngitis	345	211	344	0,07469
preventan	173	141	173	0,07469
jun	181	133	181	0,07464
/tbl	157	119	157	0,07462
panadol/nurofen	167	123	167	0,07462
větrat	184	129	184	0,07462
srdce bpn	1349	788	1349	0,07171
dýchání zostřený	640	483	640	0,0716
t 0	1 386	706	1 386	0,07134
hrdlo prosáklý	1 635	775	1 635	0,07124
prosáklý rudý	551	356	551	0,07121
prosáklý dýchání	881	463	881	0,07113
hlenový rýma	292	231	292	0,07113
s přenesený	342	251	342	0,07112
závěr rhinopharyngitis	253	199	253	0,07106
přenesený fenomenon	316	221	316	0,07104

Tabuľka 3.9: Atribúty s najvyšším IG pre diagnózu J00. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.

30gtt	21	21	21	4,68606
vincetka	6	6	6	4,68606
quatro	74	68	74	4,56407
mucopurul	17	15	17	4,50549
preventan	173	141	173	4,39098
kyselý	116	93	116	4,36724
enantem	10	8	10	4,36413
dnůnecvičit	5	4	5	4,36413
x1tbl	5	4	4	4,36413
/tbl	157	119	157	4,28626
x 30gtt	20	20	20	4,68606
zostřený hnisavý	12	12	12	4,68606
rudý hlenový	12	12	12	4,68606
quatro vincentka	11	11	11	4,68606
vincentka ibalgin	10	10	10	4,68606
nk ibalgin	10	10	10	4,68606
nk ambrosan	10	10	10	4,68606
čistý vycházka	9	9	9	4,68606
nk mucosolvan	8	8	8	4,68606
rýma smět	7	7	7	4,68606

Tabuľka 3.10: Atribúty s najvyšším PMI pre diagnózu J00. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.



Obr. 3.4: Najlepšie filtrovanie atribútov k diagóze

K30

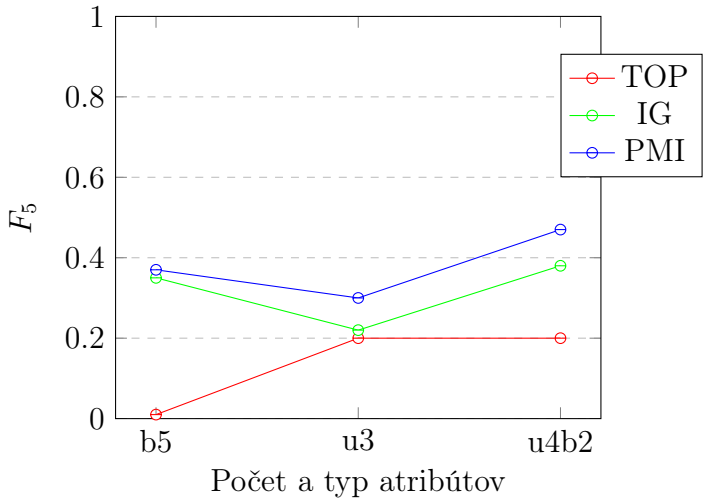
Uvedieme zoznam unigramov a bigramov s najvyššími IG a PMI pre diagnózu K30.

Atribút	**Freq**	**Dg**	**Správy**	**IG**
fob	223	82	223	-0,07191
anál	20	13	20	-0,07192
ipp	13	9	13	-0,07192
koloskopii	16	8	16	-0,07193
fob-	8	5	8	-0,07193
-pozitivní	6	4	6	-0,07194
nafouklý	15	7	15	-0,07194
meteoristicky	7	4	7	-0,07194
melenu	5	3	5	-0,07194
gfs	61	21	61	-0,07194
objektivní fob	121	68	121	-0,07542
subjektivní založit	265	108	265	-0,07545
vyš krev	108	54	108	-0,07546
fob -	80	44	80	-0,07546
biochem vyš	63	33	63	-0,07549
založit biochem	46	26	46	-0,07549
krev a	141	55	141	-0,0755
založit labor	16	13	16	-0,0755
založit lab	23	15	23	-0,07551
anál výtěr	17	12	18	-0,07551

Tabuľka 3.11: Atribúty s najvyšším IG pre diagnózu K30. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.

ipp	13	9	13	6,00414
-pozitivní	6	4	6	5,94969
anál	20	13	20	5,91316
fob-	8	5	8	5,85658
melenu	5	3	5	5,79769
meteoristicky	7	4	7	5,7273
koloskopii	16	8	16	5,53465
antrální	6	3	6	5,53465
nafouklý	15	7	15	5,43512
-krev	9	4	9	5,36473
subjektivní elev	4	4	4	6,53465
založit labor	16	13	16	6,23509
gastroent acidum	5	4	5	6,21272
subjektivní koloskopie	5	4	5	6,21272
pozitivní odběr	5	4	5	6,21272
subjektivní anál	8	6	8	6,11961
stolice 0	4	3	4	6,11961
být slyšitelný	4	3	4	6,11961
antrální gastropatie	4	3	4	6,11961
fob -pozitivní	4	3	4	6,11961

Tabuľka 3.12: Atribúty s najvyšším PMI pre diagnózu K30. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.



Obr. 3.5: Najlepšie filtrovanie atribútov k diagóze K30

Z001

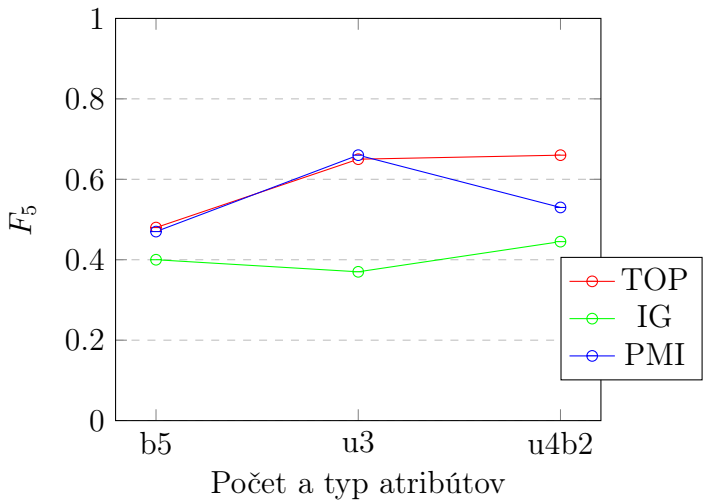
Uvedieme zoznam unigramov a bigramov s najvyššími IG a PMI pre diagnózu Z001.

Atribút	**Freq**	**Dg**	**Správy**	**IG**
dutina	4625	2851	4625	0,09549
patologie	3517	2410	3517	0,09532
orgán	2586	2044	2586	0,09528
fyziologický	2364	1860	2364	0,09485
zvětšit	3644	2313	3644	0,0947
vada	2079	1630	2079	0,09434
vrozený	1902	1556	1902	0,09434
vývoj	2061	1603	2061	0,09426
ústní	1877	1321	1877	0,09336
lymfatický	1212	1002	1212	0,09315
genitál	1837	1254	1837	0,09315
ne	3490	1869	3490	0,09312
končetina	1579	1127	1579	0,09304
patologický	1484	1038	1484	0,09283
věk	1323	970	1323	0,09281
bez patologie	3209	2409	3209	0,0922
vrozený vada	1862	1546	1862	0,09069
fyziologický nález	1497	1236	1497	0,08997
orgán dutina	1621	1280	1621	0,08993
dutina ústní	1635	1185	1635	0,08949
lymfatický uzlina	1198	1002	1198	0,08947
bez patologický	1467	1034	1467	0,08913
štítný žláza	1402	996	1402	0,08908
srdeční pravidelný	1009	829	1009	0,08905
akce srdeční	1075	839	1075	0,08897
závěr zdravý	893	750	893	0,08892
celkový zdravotní	941	763	941	0,08888
vada ne	947	764	947	0,08888
ne kýla	947	764	947	0,08888
pohlavní vývoj	943	762	943	0,08887

Tabuľka 3.13: Atribúty s najvyšším IG pre diagnózu Z001. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.

lanugo	24	24	24	4,55248
novorozenecký	23	23	23	4,55248
válková	23	23	23	4,55248
hepar1	20	20	20	4,55248
arteriales	20	20	20	4,55248
snaživě	20	20	20	4,55248
norozenecká	17	17	17	4,55248
percentil/	17	17	17	4,55248
3/3	16	16	16	4,55248
matopatie	15	15	15	4,55248
zdravý soma	133	133	133	4,55248
závěr ré	118	118	118	4,55248
vybavit dutina	45	45	45	4,55248
hmatný končetina	38	38	38	4,55248
tlouci kostka	38	38	38	4,55248
slabikovat tlouci	34	34	34	4,55248
úchop tři‘3	34	34	34	4,55248
tři‘3 prst	34	34	34	4,55248
smíšený objektivní	32	32	32	4,55248
prst sociální	32	32	32	4,55248

Tabuľka 3.14: Atribúty s najvyšším PMI pre diagnózu Z001. **Freq** - frekvencia atribútu, **Dg** - frekvencia atribútu s diagnózov, **Správy** - počet správ s daným atribútom.



Obr. 3.6: Najlepšie filtrovanie atribútov k diagóze Z001

3.5.4 3. fáza – testovanie

I10 - Človek verzus stroj

Náhodne sme vybrali 50 správ s kódom diagnózy I10, 50 správ s iným a navzájom sme ich premiešali. Potom sme požiadali lekárov o priradenie alebo nepriradenie kódu I10 ku každej z 100 lekárskych správ. Nakoniec sme ich porovnali s klasifikačným modelom natrénovaných na I10 so 400 unigramami a 200 bigramami.

Výsledky sú v tabuľke 3.15.

ID	TN	FP	FN	TP	P	R	F1	F5	ROC	K
Lekar1	46	6	19	29	0,83	0,60	0,70	0,61	–	0,50
Lekar2	50	2	32	16	0,89	0,33	0,48	0,34	–	0,32
Lekar3	49	2	29	20	0,91	0,41	0,56	0,42	–	0,38
Stroj IG	45	4	32	18	0,82	0,36	0,50	0,37	0,69	0,26
Stroj TOP	17	32	6	44	0,58	0,88	0,70	0,86	0,68	0,22

Tabuľka 3.15: Confusion matrix a metriky výkonu pre I10. **P** je Precision a **R** Recall.

Ako je vidieť z tabuľky, lekári sa sústredili hlavne na presnosť a I10 priradili správam, ktoré obsahovali zvýšené hodnoty merania krvného tlaku. V 2 prípadoch aj stroj IG aj lekár priradili I10 ale v skutočnosti sa jednalo o vedlajšie diagnózy k hlavným Z000 – Celkové lékařské vyšetření (prohlídka) a I259 – Chronická ischemická choroba srdeční NS. V žiadnom prípade sa nestalo, aby stroj IG priradil diagnózu nesprávne a zároveň ju nepriradil nesprávne aspoň jeden lekár. Stroj IG nepriradil I10 správam, ktorým I10 priradili lekári pravdepodobne preto, lebo nemal v atribútoch konkrétne číslo merania tlaku. Stroj TOP kvôli trénovaniu s cenou za chybu nastavenú v prospech Recallu klasifikoval a v 4 prípadoch neklasifikoval I10 vtedy, keď ju Stroj IG klasifikoval. Celkovo môžeme povedať, že Stroj IG obstál v porovnaní s človekom veľmi dobre. Výsledky pre každú správu ako aj samotné správy sú na priloženom CD.

Poznámka. Lekar1 bol práve ten lekár, ktorý nás upozornil na prípadnu súvislosť I10 a predpisovania liekov.

Testovacie výsledky

DG	TN	FP	FN	TP	P	R	F1	F5	ROC	K
I10	439 541	30 968	26 435	26 862	0,46	0,50	0,48	0,50	0,78	0,42
H660	515 544	3 922	1 130	3 210	0,45	0,74	0,56	0,72	0,95	0,56
J00	503 877	6 710	9 823	3 396	0,34	0,26	0,29	0,26	0,77	0,28
K30	514 806	3 696	4 587	717	0,16	0,14	0,15	0,14	0,64	0,14
Z001	501 197	2 407	16 057	4 145	0,63	0,21	0,31	0,21	0,64	0,12

Tabuľka 3.16: Výsledky klasifikačného modelu trénovaného algoritmom Naive Bayes. Atribúty boli zvolené informačným ziskom (IG). **P** je Precision a **R** Recall.

3.6 Zrhnutie

Pri sumárnych výsledkoch na počte správ 100 000 a viac sa predspracovanie ich textu vo výsledkoch nemusí výrazne prejaviť. Ukázalo sa taktiež, že od 300 unigramov a viac sú výsledky pomerne stabilizované. Pri výkone bežného počítača nám prišlo 600 atribútov ako únosný počet pre trénovanie. Najfrekventovanejšie atribúty zaručia najvyššiu metriku F_5. Z použitých klasifikačných algoritmov výsledkami ako aj nízkou výpočtovou náročnosťou najlepšie vyšiel Naive Bayes. Pri

meraní výkonu pomocou F_5 je vhodne použiť Meta Cost s Naive Bayes a cenou päť krát vyššou za nepriradenie kódu diagnózy ako za nesprávne priradenie.

Avšak nám záleží na každej jednej správe. Pre zanedbanie niektorých chybne klasifikovaných správ s tým, že globálne štatistiky to neovplyvní, nie je miesto. Preto odstránenie nealfanumerických znakov, zachovanie niektorých tokenov ako merania tlaku a dávkovania a aplikovanie morfológie sú veľmi žiadúce techniky predspracovania. Bigramy môžu odhaliť skryté súvislosti, preto sme používali kombináciu 400 unigramov s 200 bigramami. Pri výbere atribútov sme mali možnosti vybrať tie s najvyšším výskytom, vzájomnou informáciou (PMI) a informačným ziskom (IG). Atribúty s najvyšším výskytom neberú do úvahu diagnózu, PMI berú do úvahy iba diagnózu a snažia sa vyberať atribúty, ktoré sa pri iných diagnózach takmer nevyskytujú. IG dosahuje najlepších výsledkov, lebo berie do úvahy entropiu celých dát a entropiu atribútu.

Najpočetnejšie atribúty nám ukázali, že F_5 je skôr pomocná metrika a vhodná metrika je F – *Measure* za doprovodu *Kappa* a *ROC*.

Z porovnania stroj vs človek na priradenie kódu diagnózy I10 sme zistili, že filter IG klasifikuje podobne ako lekári. Zároveň to pozorovanie ukazuje dôležitosť relevancie atribútov a prázdny význam čísel, čo je spôsobené kontextom správ. Hoci si na testovacích dátach vedie H660 podobne dobre ako I10, ale bez porovnania s lekárom nemôžeme s istotou prehlásiť, že automatické priraďovanie H660 funguje a nie je iba špecifické pre konkrétne dáta.[5]

3.7 Budúca práca

Niektoré atribúty určujú kód diagnózy jednoznačne, a práve tým by bolo vhodné priradiť nejaké váhy. Ďalej vziať do úvahy negatívne slová ako *nemá*, *nenalezen*, *žádný*,*alergie:0* a pod. Užitočné by bolo aj nahradenie numerických hodnôt za ich sémantický význam. Napríklad pri meraní tlaku namiesto čísel by bolo TK 120/80 TKOK, TK 135/85 TKHIGH a podobne.

Ponúka sa veľký priestor pre klasifikáciu voči viacero diagnózam súčasne (*Multiclass classification*).

Čo sa týka dát, boli by veľmi prínosné správy s vedľajšími diagnózami a minulý kontext, t. j. históriu.

[5]Pri I10 to môžeme prehlásiť preto, že aj nelekárskym pohľadom vidíme, že IG vyberá atribúty určujúce I10 všeobecne. U H660 to aspoň autor posúdiť naozaj nevie.

Záver

Vyvinuli sme vlastný systém schopný klasifikácie textu. Medzi jeho hlavné možnosti patrí predspracovanie textu, selekcia atribútov a zber výsledkov. Na samotnú klasifikáciu sme použili externú sadu algoritmov programu WEKA. Na trénovanie sme použili 100 000 správ a na testovanie 500 000.

So zameraním na priraďovanie kódu diagnózy I10 sme natrénovali model. Na testovacej vzorke 100 správ, ktoré boli pre porovnanie binárne ohodnotené lekármi obstál model veľmi dobre. Dokonca lepšie, ako jeden z lekárov.

Zároveň, pre hlbokú odbornosť správ, súčasťou vyhodnotenia výkonnosti programu pre konkrétny kód diagnózy musí byť aj priame porovnanie s lekármi. Týmto sa zistí, či bol alebo nebol program natrénovaný na špecifickú sadu dát, a či zvolené atribúty určujú kód danej diagnózy všeobecne.

Myslíme si, že táto práca môže poslúžiť ako vhodný základ jej rozšírenia uvedené v 3.7.

Zoznam použitej literatúry

[1] Witten, Ian H., Frank, Eibe a Hall, Mark A.. *Data Mining Practical Machine Learning Tools and Techniques*. 3. vydanie. Burlington: Morgan Kaufmann, 2011. ISBN 978-0-12-374856-0.

[2] Manning, Christopher D. a Schuetze, Henrich. *Foundations of Statistical Natural Language Processing*. 1. vydanie. Massachusetts: The MIT Press, 1999. ISBN 0-262-13360-l.

[3] Ng, Andrew. *Machine Learning*. [online kurz]. In CS 229 Machine Learning Autumn 2008. Záznam dostupný z <http://www.youtube.com/watch?v=UzxYlbK2c7E>.

[4] Aggarwal, Charu C. a Zhai, ChengXiang. *A Survey of Text Classification Algorithms*. Kapitola v Mining Text Data. Ed. Charu C. Aggarwal a ChengXiang Zhiang. 1. vydanie. New York: Springer US, 2012, s. 163 – 222. ISBN 978-1-4614-3222-7.

[5] Domingos, Pedro. *MetaCost: A General Method for Making Classifiers Cost-Sensitive*. In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, 1999, s. 155–164. ACM Press.

[6] Hardikar, Surbhi, Shrivastava, Ankur a Choudhary, Vijay. *Comparison between ID3 and C4.5 in Contrast to IDS*. In VSRD- Internation Journal of Computer Science & Information Technology, 2012, s. 659 –667. ISSN 2231-2471. Dostupné na internete: <http://www.vrsdjournals.com>.

[7] Raab, Jan. *Morče - Czech Morphological Tagger*[počítačový program]. Ústav formální a aplikované lingvistiky, Univerzita Karlova v Praze. Dostupné z <http://ufal.mff.cuni.cz/morce/index.php>

[8] Dolamic, Ljiljana . *CzechStemmerAgressive*[zdrojový kód počítačového programu]. University of Neuchatel, Switzerland. Dostupné z <http://members.unine.ch/jacques.savoy/clef/CzechStemmerAgressive.txt>

[9] Neznámy. *Stopwords list in Czech*[textový súbor]. University of Neuchatel, Switzerland. Dostupné z <http://members.unine.ch/jacques.savoy/clef/czechST.txt>

[10] Machine Learning Group at the University of Waikato. *Weka 3: Data Mining Software in Java*[počítačový program]. Verzia 3.6.10 [cit. 2013-12-05]. Dostupné z <http://www.cs.waikato.ac.nz/ml/weka/>

[11] Bouckaert, Remco R. *WEKA Manual for Version 3-6-10*[online]. University of Waikato, 2013 [cit. 2013-12-05]. Dostupné z <http://prdownloads.sourceforge.net/weka/WekaManual-3-6-10.pdf?download>

[12] *MONO*[online]. Dostupné z <http://www.mono-project.com/Main_Page>

[13] Microsoft Corporation. *Microsoft .NET Framework 4.5*[počítačový program]. Dostupné z <http://msdn.microsoft.com/en-us/library/w0x726c2(v=vs.110).aspx>

[14] NUnit.org. *NUnit*[počítačový program]. Dostupné z <http://www.nunit.org/index.php?p=home>

[15] *Command Line Parser Library*[počítačový program]. Dostupné z <https://commandline.codeplex.com/>

[16] Breiman, Leo a Cutler, Adele. *Random Forests*[online]. University of Waikato, 2013 [cit. 2013-12-05]. Dostupné z <http://www.stat.berkeley.edu/ breiman/RandomForests/>

[17] Zvára, Karel a Kašpar, V. *Identifikace jednotek a dalších termínů v českých lékařských zprávách.* Centrum biomedicínské informatiky, Ústav informatiky AV ČR [cit. 2013-12-05]. ISSN 1801-5603. Dostupné na internete: <http://www.ejbi.org/en/ejbi/article/61-cs-identifikace-jednotek-a-dalsich-terminu-v-ceskych-lekarskych-zpravach.html>.

[18] Prečková, Petra. *Jazyk českých lékařských zpráv a klasifikační systémy v medicíně.* Centrum biomedicínské informatiky, Ústav informatiky AV ČR [cit. 2013-12-05]. ISSN 1801-5603. Dostupné na internete: <http://www.ejbi.org/en/ejbi/article/53-cs-jazyk-ceskych-lekarskych-zprav-a-klasifikacni-systemy-v-medicine-.html>.

[19] Zhang, Harry. *The Optimality of Naive Bayes.* In Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004), 2004. AAAI Press.

[20] TAPE, Thomas G. *Interpreting Diagnostic Tests.* In Darwin, University of Nebraska Medical Center, [cit. 2013-12-1]. Dostupné na internete: <http://gim.unmc.edu/dxtests/Default.htm>.

[21] Zvára, Karel a Štěpán, Josef. *Pravděpodobnost a Matematická Statistika.* 4. vydanie. Praha: MATFYZPRESS, 2006. ISBN 80-86732-71-7.

[22] *Mezinárodní klasifikace nemocí.* 10. revízia, aktuálizovaná verzia k 1.1.2013 [cit. 2013-12-01]. Dostupné na internete: <http://www.uzis.cz/cz/mkn/index.html>.

[23] *Ústav zdravotnických informací a statistiky ČR.* Číselník diagnóz LPZ k 1.1.2013 [cit. 2013-12-01]. Dostupné na internete: http://www.uzis.cz/system/files/dokumenty/DGLPZ.xml>.

[24] World Health Organization. *History of the development of the ICD.* [cit. 2013-12-04]. Dostupné na internete: <http://www.who.int/classifications/icd/en/HistoryOfICD.pdf>.

Prílohy

A. Dokumentácia k ARFFBuilder

ARFFBuilder je počítačový program naprogramovaný pre účely tejto práce. Obsahuje konzolový program ARFFBuilder.Core, ktorý je plne kompatibilný s multiplatformným projektom *MONO*[12] a *ARFFBuilder.GUI*, čo je grafická užívateľská nadstavba podporovanú iba pre platformu Microsoft .NET 4.5[13]. *ARFFBuilder.Core.NUnit* je sada unit testov pre *ARFFBuilder.Core* postavených na knižnici *NUnit*[14]. Súčasťou riešenia je ešte vlastný program *WEKAResultsToCSV*, ktorý pre danú diagnózu hromande spracuje výsledky s WEKA a uloží ich do CSV súboru.

lekárskych správ a k nim priradených diagnóz. Vygenerovaný súbor je vhodný ako vstupný súbor pre Weka[10].

ARFFBuilder.Core - Užívateľská dokumentácia

Program *ARFFBuilder.Core* zo vstupného súboru generuje súbor vo formáte *ARFF*, ktorý dokáže použiť napríklad nami používaný externý program *WEKA*. *WEKA* je súhrn impletácií rôznych algoritmov a technik strojového učenia [11][10].

Vstupný súbor je ľubovoľný textový súbor obsahujúci na začiatku každého riadku klasifikačnú triedu a text, ktorý je priradený danej klasifikačnej triede. Trieda a text sú od seba oddelené jednou medzerou.

V prípade nezvolenia nijakého filtrovania atribútov, program ich vyberie podľa frekvencie výskytu. Po zadaní správnych parametrov, viď 3.7, program vypíše na koznolu *I'm working ...* a po skončení výpočtov *I'm done.* a ukončí sa. Pri zadaní nesprávneho parametru sa vypíse nápoveda a program sa ukončí.

parameter	popis
--help	Vypíše nápovedu a ukončí sa.
-i F	*F* vstupný súbor.
-o F	*F* ARFF súboru.
-d T	*T* je názov diagnózy na trénovanie.
-f F	*F* ARFF súboru v kódovaní UTF-8 pre náčítanie atribútov.
-b N	*N* je maximálny počet bigramov.
-u N	*N* je maximálny počet unigramov.
--uf=N	Unigramy s nižšiou frekvenciou ako *N* ignoruje.
--bf=N	Bigramy s nižšiou frekvenciou ako *N* ignoruje.
--pmi	Vyberie atribúty s navyšším PMI.
--ig	Vyberie atribúty s navyšším IG.
-m F	*F* s morfológickym slovníkom vo formáte CSTS.
-s F	*F* so zoznamom stopwords. Slvo na riadok.
--idf	Vygeneruje stopwords na základe IDF
--idf-count=N	*N* je maximálny počet vygenerovaných. stopwords.
--idf-output=F	*F* pre vygenerované stopwords pomocou IDF.
--stemmer	Použije vstavaný stemmer na českú gramatiku.
--keep-punctuation	Ponechá nealfanumerické znaky.

F predstavuje cestu k súboru, *T* text a *N* celé číslo.

Tabuľka 3.17: Zoznam vstupných parametrov pre ARFFBuilder.Core

Použitie parametru *--idf* spôsobí ignorovanie prípadného parametru parameter *-s F*.

ARFFBuilder.GUI - Užívateľská dokumentácia

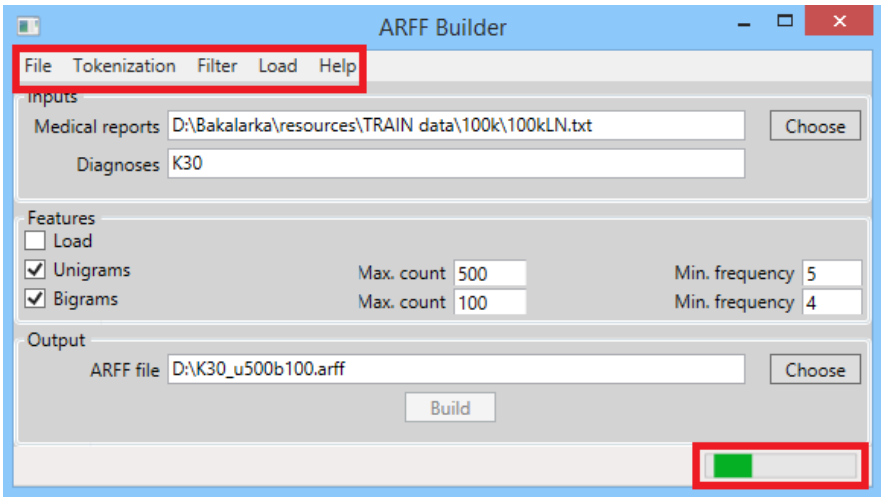
ARFFBuilder.GUI je grafické užívateľské rozhranie pod Windows s .NET Frameworkom 4.5 a vyšším nad programom ARFFBuilder.

Hlavné okno sa skladá z hlavného menu, stredu a progress baru. Hlavné menu a progress bar sú na obrázku 3.7 je vyznačené červeným obdlžníkom. Ako je ďalej na obrázku vidno, v strede je možno zvoliť klasifikačný súbor, diagnózu, typy atribútov a výstupný súbor vo formáte ARFF. Možnosti hlavného menu popisuje nasledovný zoznam podľa obrázka

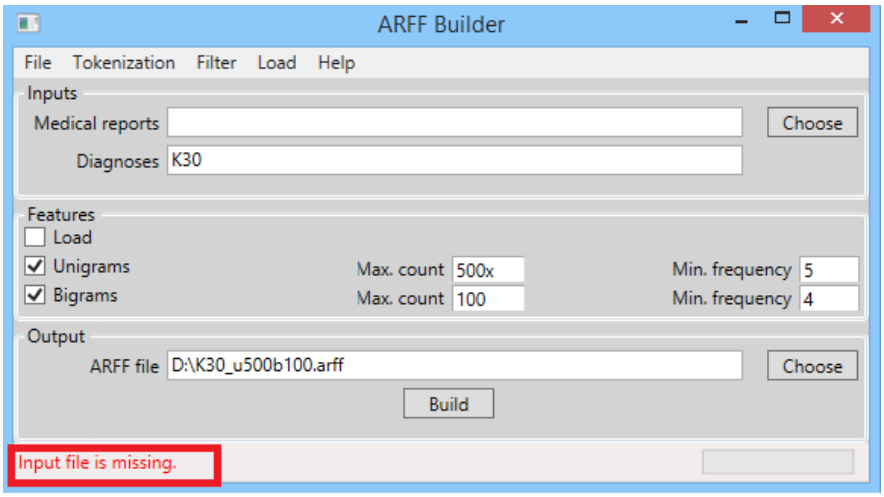
- File - zavrieť súbor,
- Tokenizaiton - použiť stemmer, ignorovať stop words, zachovať nealfanumerické znaky,
- Filter - filtrovať atribúty podľa IG alebo PMI. IDF vygeneruje stopwords,
- Load - načítať stop words, atribúty alebo morfológiu vo formáte CSTS,
- Help - zobraziť dokumentáciu alebo info o programe.

V prípade, že všetky povinné políčka sú vyplnené a žiadne nastavenie neobsahuje syntaktické chyby, po kliknutí na tlačítko *Build* sa spustí výpočet. Inak

sa zobrazí na ľavo od progress baru chybová hláška, ako je vyznačené na 3.8. Priebeh výpočtu signalizuje progress bar ako aj neaktívne tlačítko *Build*.



Obr. 3.7: Hlavné okno ARFFBuilder.GUI



Obr. 3.8: Hlavné okno ARFFBuilder.GUI

Hromadné spušťanie výpočtov

Pre skúšanie rôzných klasifikačných tried, tokenizácie, atribútov a klasifikátorov sme napísali skripty. Sú k dispozícií na priloženom CD v adresári *scripty*. Podadresár *bat* obsahuje scripty pre príkazový riadok vo Windowse, resp. podadresár *sh* pre shellovský interpreter.

Správe fungovanie scriptov vyžaduje program *ARFFBuilder.Core* a *WEKA*. *WEKA* aplikuje klasifikátory, ktoré sa spúšťajú zo *classifiers.bat* resp. *classifiers.sh*. Cesty k programom a adresáre sa nastavujú v súbor *paths.bat* resp. *paths.sh*. Scripty vedia pracovať s adresárov štruktúrou zhodnou s tou, na priloženom CD.

Adresárová štruktúra

Hromadný výpočet predpokladá klasifikáciu viac ako jednej triedy s rôznymi typmi a počtami parametrov, skúšanie rôznej tokenázacie, filtrovania atribútov, stopwords atď. Preto sú skripty predpripravené na adresárovú štrukturu symbolicky zapísanú T/A/O, kde T predstavuje klasifikačnú triedu, A typ a počet atribútov a O iné nastavenie. V prípade tejto práce sme použili konkrétne I10/u500b100/filter, teda pre diagnózu I10, 500 unigramov a 100 bigramov a filter atribútov, v našom prípade PMI a IG.

Názvy nie sú pevné dané a majú najmä slúžiť pre orientáciu, aby dalo jednoducho určiť, aké nastavenia boli použité pre daný výsledok.

Nastavenie ciest a hodnôť premenných

Cesty k súborom ako aj hodnoty iných premmenných používaných ďalej v skriptoch sa nastavujú v súbore paths.bat resp. path.sh. Uveďme ich zoznam s popisom.

- ROOT - cesta ku koreňovému adresáru. Odporúčame sa máť všetky ďalšie súbory v ňom.
- ARFFBuilder - cesta k programu ARFFBuilder.Core.
- RECORDS - ceska k vstupnému súboru dát, z ktoré budu charakterizované ARFF súborom.
- STOPWORDS - ceska k zoznamu stopwords.
- MORPHOLOGY - cesta k morfologickému slovníku vo formáte CSTS.
- WEKA_CP - ceska k súboru weka.jar, tzv. class path.
- WEKA_MEMORY - veľkosť pamäte, ktorú môže WEKA alokovať.

Nastavenie názvu výstupných súborov

Ako názov výstupných súborov sa použili diagnóza, suffix predstavujúci nastavenia program *ARFFBuilder.Core* a použitý klasifikátor. Klasifikátor sa nastavuje v premennej *CLASS*, zvyšok v jednotlivých scriptoch v podadresári pre iné nastavenia v premennách *DG* a *SUF*, symbolicky vyššie značeného *O*.

Spustenie výpočtu

Celý výpočet sa pustí skriptom v poslednom adresári adresárovej štruktúry popísanej vyššie, t.j. adresár symbolicky označenom *O* v príkazom riadku resp. interpretri shellu. Pre univerzárnosť berie ako prvý parameter názov diagnózy.

Hromadné získanie výsledkov

Pri hľadaní najlepšiej kombinácie techník strojového učenia je potrebné veľa výsledkov pre porovnanie. Scripty ako výsledok generujú textové súbory .txt a binárne súbory .model. Z .txt súboru nás hlavne zaujíma *TP*, *FP*, *Recall*, *Presision*, *F-Measure*, *ROC aréna*. Progam navyše pridá ešte dva stĺpce. Prvý, identifikátor, sa výrobí konkatenaciou názvu adresára a názvu daného súboru. Druhý, kontrola, slúži na overenie fuknčnosti programu, t.j. že vyextrahoval správny riadok yes. Pokiaľ je v poslednom stĺpci iná hodnata ako yes, jedná sa o chybu extrahovania z daného súboru.

Súbory .model slúžia ako klasifikačný model pre program WEKA pre testovanie testovacej sady dát.

WEKAResultsToCSV - užívateľská dokumentácia

Program WEKAResultsToCSV berie ako vstupy .txt súbory, ktoré vygenerovala WEKA binárnou klasifikácou na nejakú klasifikačnú triedu. Z nich extrahuje pre výsledok do jedného .csv súboru vhodného pre súmarné zobrazenie v nejakom tabuľkovom procesore.

Jeden výsledok v závere obsahuje identifikátor, *TP*, *FP*, *Recall*, *Precision*, *F-Measure*, *ROC aréna* a *Kontrolu*.

ARFFBuilder.Core - Programátorská dokumentácia

Program je napísaný v jazyku *C#* využitím štýlu a techník *objektovo-orientovaného programovania*. V tejto programátorskej dokumencii sa sustredíme na celkové rozvrhnutie programu a všeobecný popis jeho jednotlivých logických časti na úrovni tried. Pre konkretný popis jednolivých metód a členov triedy viď komentovaný zdrojový kód na priloženom CD.

Entity

Entity sú definované v samomstatnom namespace *ARFFBuilder.Core.Entity*. Sú logicky rozdelené na atribúty, kolekcie atribútov, nastavenie a inštancie.

Atribúty

Ako entity atribútov môžeme chápať triedy uvedené naledujúcim zoznamom.

- *FeatureEnum* - Enum typov atribútov. Má hodnoty Unigram a Bigram. Slúži najmä pre čitateľnosť a bezpečnosť pri práci s typmi.
- *Unigram, Bigram* - Potomkova abstraktnej triedy triedy FAttribute. Definujú vlastné konštruktory.
- *FAttribute* - Definuje všeobecné vlastnosti atribútu ako jeho meno, frekvenciu, PMI, IG, počet textových správ s trénovanou triedou a počet všetkych

inštancii, v ktorých sa vysktyuje. Posledné dve vlastnosti sa využívajú najmá pri počítaní PMI a IG. Ďalej definuje rovnosť dvoch atribútov a statické metódu pre zistenie typu atribútu a počet jeho výskytov v texte.

Kolekcie atribútov

Triedy *UnigramCollection* a *BigramCollection* sú potomkami triedy *AttributeCollection* a definujú metódu jej jedinú abstraktnú metódu TextMsgParse slúžiacu na parsovanie atribútov daného typu z textu. Ďalej abstrakntá trieda *AttributeCollection* okrem klasických metód kolekcií ako pridanie a získanie všetkých atribútov umožnuje vrátiť atribúty v poradí podľa nastavenia filtra. Teda podľa frekvencie, PMI alebo IG.

Nastavenie a inštancie

Trieda *Settings* ukláda nastavenia zvolené užívateľom ovplyňújúce správanie programu cez mnohé triedy. Preto existuje abstraktná trieda SettingsClass nesúca v sebe inštanciu triedy Settings, od ktorej dedia všetky triedy, ktoré menia svoje správanie podľa užívateľkých nastavení. Členy triedy Settings sú takmer zhodné s tým, čo všetko môte nastaviť užívateľ.

Trieda *Instance* definuje štruktúru jednej inštancie, t. j. dvojica klasifikáčná trieda a textová správa. Trieda *Records* načítava inštancie so vstupných dát.

Budovanie atribútov a tokenizácia

Za budovanie atribútov je zodpovedný interface *IFeatures* s dvomi metódami *BuildFeatures* a *BuildAllUnigrams*. Prvá metóda vybuduje atribúty podľa nastavení. Druha metóda je postavená na tom, že slovo, t.j. unigram, je najmešia jednotka v strojovom učení a NLP, s ktorou má zmysel pracovať. Využíva sa pri tokenizácií a výrobe stopwords pomocou IDF. Potomkom *IFeatures* je trieda *Features*, ktorá obsahuje privátnu metódu na načítanie atribútov z ARFF súboru. Jej použitie samozrejme závisí od nastavení, ktoré sa predávajú, preto nepotrebuje byť v interface.

Tokenizácia má svoj namespace *ARFFBUilder.Core.Tokenization*. Volá sa cez interface *ITokenizer* , ktorý má dve metódy. *Tokenize* tokenizuje celý text a *TokenizeWord* jedno slovo. Tokenizer interne využíva inštancie interfaceov *IMorphology*, *IStemmer* a *IStopWords*, ktoré sú zodpovedné za prácu s morfógiou, stemmerom a stop words.

Výpočty a iné utility

Všetky triedy v tejto časti sú statické lebo nezávisia na vnútornom stave objektu ale pre rovnaký vstup dávajú vždy ten rovnaký výstup. Trieda *Computations* počíta IDF, PMI, a IG. Trieda *SringUtils* pracuje s reťazcami a dokáže odstráňovať intepunkciu alebo úvodzovky. Posledná čisto statická trieda je *InstanceUtils*, ktorá dokáže parsovať text na všetky typy atribútov.

Vstup a výstup

Program v triede *MainClass* obsahuje metódu *main*, ktorá sa volá ako prvá po štarte. Tá naparsuje vstupné parametre pomocou knižnice *CommandLine Parser Library*[15] a predá riadenie inštancii IBitmap. Interface *IBitmap* obsahuje dve metódy. *BuildBitmap* vybuduje bitmapu zo vstupného súboru a *WriteToARFF-File* vypíše vybrané atribúty a vybudovanú bitmapu do ARFF súboru na cestu zadanú užívateľom.

Rozšírenia programu

V tejto časti ukážeme ako bezproblémovo rozširiť program.

Nový typ atribútu

Ako vzor si vziať triedu *Unigram* alebo *Bigram* a prácu s ňou. Je potrebné vyrobiť a upraviť nasledujúce triedy.

- *FeaturesEnum* - pridať nový typ
- *NovýTyp* a *NovýTypColleciton* - vyrobiť nový triedy predstavujúce typ a jeho kolekciu.
- *Features* - nová privátna metóda *InitNovýTyp* a použiť ju vo verejnej metóde *BuildFeatures*.
- *ConsoleOptions* - sprístupniť ho užívateľovi a nastaviť v *Settings.Features*.

Nový filter pre atribúty

Je potrebné upraviť nasledujúce triedy.

- *Computations* - pridať statickú metódu pre výpočet nového filtru.
- *FAtrribute* - pridať ako property.
- *Settings* - pridať ako verejného člena a pridať do metódy *FilterDescription*.
- *AttributeCollection* - pridať to privátnej metódy. *OrderAttributeBySettings* a verejnej *ApplyFilter*.
- *ConsoleOptions* - sprístupniť ho užívateľovi a nastaviť v *Settings.Features*.

Tokenizácia

V prípade pridania nového vstavaného stemmera alebo nového formátu morfológie stači vyrobiť novú triedu podedenú od *IStemmer* alebo *IMorphology*. Inak by bolo slušné vytvoriť nový interface a jeho inštanciu v namespace *ARFFBuilder.Core.Tokenization*. Na sprístupnenie užívateľovi upraviť triedu *ConsoleOptions* a v podobnom duchu rozšíriť vnútornú triedu *Settings.TokenizerUsage*. Nakoniec je potrebné prepojiť nastavenia v konštruktore a metóde *TokenizerWord* v triede *Tokenizer*. V prípade nového druhu tokenizácie vytvoriť privátny člen v triede *Tokenizer*.

B. Obsah CD

CD obsahuje tisíc lekárskych správ, ARFFBuilder, skripty hromadného spustenia a výsledky jednotlivých fáz vypočtu.

Pre učelý tejto práce firma TERSINIDA a.s. prístupnila tisíc lekárských správ na tomto CD.